JET
CONFERENCE

QUALITY and VALIDITY

in INSET EVALUATIONS

CONFERENCE PROGRAMME

EVALUATION AND ACCOUNTABILITY
Dr Nick Taylor, JET

Over the past four years The Joint Education Trust (JET) has disbursed a total of R56,8 million to 98 NGO programmes working in the field of Teacher Development and Support. Our projections are that in the next three years the Trust will spend an equivalent sum in this sector.

Information on the activities of NGOs is difficult to collect systematically, but our best estimates indicate that JET supports 70% of programmes working in the teacher development sector, and provides one third of all funding to these NGOs. We calculate that JET grantees directly reach almost 38 000 teachers, which is over 10% of the total teaching force.

Operations of this magnitude require some form of evaluation and we would be neglecting our obligations to our donors, the NGO grantees, the teachers involved in these programmes, the broader educational community, and the general public if we made no attempt to ascertain the effects of this activity and to subject the processes and results of our evaluations to public debate. Today's conference is the first step in this direction.

Evaluation is essentially about accountability. Three closely related issues are involved in using evaluation as a tool for accountability:
• donor accountability - ensuring that donors are informed about the effects of their donations.
• scientific accountability - basing evaluations on the accepted canons of science.
• systemic accountability - using the results to improve the effectiveness of programmes evaluated and relevant parts of the education system.

These three issues involve questions of public interest which include, but also go way beyond the concerns of individual NGOs and their beneficiaries. It is appropriate, therefore, that in his contribution to today's conference Peter Weingart discusses the current trend, in many countries and in all areas of knowledge production, towards external accountability.

Donor Accountability

On the more specific question of donor accountability it is worth reflecting on the sources of JET's funds, and the purposes these donations are intended to serve. At present all finances of the Trust derive from the business partners of JET: 18 of South Africa's largest private sector companies. There is a perception in some quarters that, since this money comes largely from pre-tax profits, it does not cost the company or its shareholders anything. This perception is incorrect. Only one half of these donations, if the Receiver of Revenue is generous, would have been paid in tax. The rest comes directly out of the pockets of the shareholders. Why would they want to do this? Partly for marketing purposes, to generate public goodwill towards their companies. And partly as an investment in human resource development: to improve the skills base of the country, and hence the profitability of business. The overriding evaluation questions for these donors are: is this investment generating quality returns in terms of improved teacher skills? Should business continue to donate to teacher development?

What about that half of these donations that would have been paid in tax? In the final analysis this money comes, albeit indirectly, out of the pockets of all citizens who pay taxes. In this regard it is

sobering to consider that for every R2 000 granted in tax concessions, the government can afford to keep one less child at school. For these donors the questions which evaluations must answer are essentially about trade-offs: is it in the public interest to facilitate donations to NGOs rather than to use the money to strengthen the public sector? Such questions, in turn, are also ultimately concerned with the quality of services provided by NGOs, and whether NGOs deliver such services more effectively than the state.

Scientific Accountability

What is educational quality? How is it measured? These questions take us into the technical nitty gritty of evaluations and the terrain of scientific accountability.

Within JET, research into these questions is directed from our Evaluation Division under the leadership of Penny Vinjevold. Over the last three years JET has commissioned 53 programme evaluations in the five sectors in which JET works. 11 of these have been in the Teacher Development sector. She is also busy directing surveys of each of our sectors. The first of these was a look at the contribution of NGOs to teacher education, as part of the National Audit of Teacher Education commissioned by the National Department of Education.

As part of that survey, we undertook an analysis of recent evaluations of NGO teacher development programmes. (See Appendix 1 and 1a). This study unearthed 54 evaluations covering 33 INSET programmes.

In the first instance, our analysis of these evaluations revealed much about the present state of evaluation practice in South Africa. All the studies, without exception, produced fascinating and useful insights into teaching and learning practices, conditions in schools and the valuable work undertaken by NGOs. However, in the majority of cases, the methods employed in the studies lie below the quality threshold required to inspire confidence in the validity of their principal findings.

For example, a number of evaluators concluded, on the strength of single class visits to a small number of project teachers, that the practices of these teachers had improved and that this improvement was due to the project intervention. Without careful comparison and control such conclusions are totally invalid. In his paper Johan Louw focuses on issues of method and, in particular, on the kinds of principles and procedures which need to be followed in order to ensure the validity of evaluation findings.

Our own survey, mentioned above, identifies the qualitative/ quantitative stand-off, which divides the field of evaluators into two distinct and bitterly opposed camps, as the single largest contributor to the present paralysis in the field. It is interesting to note that in his paper Peter Weingart characterises this stand-off as being of a largely ideological nature. Our own conclusion is that the whole spectrum of quantitative and qualitative methods needs to be used where appropriate, and that the nature of the information required should be the determining factor in choosing evaluation tools rather than some pre-determined political position.

This is not to ignore the fact that deep epistemological reefs lurk beneath all attempts to cross paradigm boundaries. These problems include differences about the nature of what it is that the evaluator is observing, how it should be observed, and how our observations should best be represented. In their respective presentations, Jonathan Jansen and Johan Muller look at some of these issues.

Systemic Accountability

It is often said that NGOs are best at piloting innovation. Opinion differs as to whether successful models should then be taken to scale by government, NGOs, or some combination of the two. Whichever way this argument is decided, the first step is to demonstrate success. It is all very well for NGOs to be highly independent in their operations, but all educational activities contribute to a single system, and the very lives of NGOs depend on their demonstrating their positive contribution to this system. Evaluation is a useful tool in

measuring the nature and extent of this contribution. It is in this sense that evaluations ensure systemic accountability. Our survey of recent evaluations conducted in the INSET sector also told us much about the systems, practices and effects of the NGO programmes under evaluation. There is an enormous amount that can be said here, but I will highlight only one thread of this rich body of information and insight: the effects of INSET programmes on teaching and learning.

Here is what one of the evaluators has to say on this issue:

"The struggle to introduce an alternative and more enlightened methodology has been won. The next target is becoming clear: to re-introduce and reinforce awareness of the need for more cognitive effort to be routinely required of pupils. It is not nearly enough for teachers to merely 'interact' with pupils, without reference to the quality of the work being demanded from the pupils."
(Peacock, 1995, 2, See Appendix 1)

"most of the children most of the time do nothing... and certainly nothing as cognitively demanding as individual reading and writing."
(ibid., 13)

This example is important for at least two reasons. The first is obvious: the teaching/ learning situation lies at the heart of the educational enterprise, and improving its quality is the goal of most INSET interventions. And Peacock's observations, widely echoed in our survey, tell us that most programmes only get halfway there. They change the forms of classroom behaviour but usually do not improve its outcomes in terms of pupils' knowledge, skills and attitudes.

At least part of the problem lies in the fact that teaching methods are seen as the end point of INSET programmes: the goal is to move from teacher centred methods, delivered predominantly through the medium of lectures, to child centred, problem-oriented methods in which the pupils must be seen to be sitting in groups and talking amongst themselves. The result is children sitting in groups and talking, but, by and large, not learning to read, write and calculate with any greater proficiency. There is a second and perhaps even more important reason why focusing on intermediate states, such as teaching forms, rather than learning outcomes is a problem. We live in an age of uncertainty in which we have lost confidence in the goals and contents of education. We feel fairly certain that we need citizens who are creative, flexible, innovative, self-confident and tolerant, but we are not sure about how to get there. Under these circumstances, the inflexible insistence on following particular processes - such as always having to work in groups - is a new kind of tyranny that once again robs teachers and pupils of any possibility of innovative spark and adaptation to changing conditions.

Case Studies

So far my introduction has been concerned with the big, and at times, rather abstract issues. To keep our noses to the practical task, our afternoon session is devoted to a case study. Here Jennifer Bisgard, Noleen Barry and Penny Vinjevold reflect the perspectives of evaluator, programme director and funder. This is where the big questions and the technical issues intersect in an actual evaluation.

Conclusion

Evaluation in South Africa is in an early stage of development. It was not possible before April 1994 to get past the ideological heat in order to address the real issues. But I believe that we will go a long way in a short time now that the apartheid fence has come down. I hope that today's conference will be remembered as a significant landmark in this debate, and I wish you all well in your deliberations.

ACADEMIC RESEARCH, INTERNAL VALIDITY AND PROGRAMME EVALUATION

Professor Johann Louw, University of Cape Town

Social programmes are, by their conception, design and implementation, directed at alleviating social problems: reducing juvenile delinquency, increasing knowledge of sexually-transmitted diseases; providing inservice training to teachers; preventing drug and alcohol abuse among teenagers; etc. It therefore is difficult to avoid the conclusion that social programmes follow an intrinsic cause-and-effect logic: "IF we enrich the teaching practices of school teachers by providing an in-service training programme, THEN the educational performance of school children taught by them will improve". In other words, without the programme, the social problems at least will continue, or may worsen; with the intervention, the situation will be improved in definable ways. Once this is recognised, it follows that the evaluation of such programmes has to include an evaluation of outcomes as well.

The specific question put to me was: Can one derive validity criteria for programme evaluation from academic research?

There are five elements required to structure a coherent answer to this question: we need to discuss programme research, programme outcomes, causality, research design, and programme theory. Hopefully, this also will illuminate the overall conference themes.

Programme research

In the turbulent world of everyday programme evaluations, with its emphasis on practical utility and timely information, the term "academic research" is often used in a somewhat derogatory way, which suggests that it is not responsive to these demands. Cordray and Lipsey (1986) have made a useful distinction in this regard, between programme research (or "academic research") and programme evaluation: "The former establishes the presence of merit within an intervention, action, or entity; the latter illuminates the mechanisms and causes of the features judged meritorious" (p. 22). Programme evaluation for these authors represents essentially a service-oriented mode of inquiry, revolving around the practical information needs of evaluation clients and their need for a credible judgement of a programme's value. In the short term, programme research has no pretensions to be responsive and useful to stakeholders. Instead it focuses on questions of how programmes actually work, on establishing the effects of programmes on a range of dependent variables, on the generality of those relationships, etc. The utility of programme research therefore is in attempts, in the longer term, to develop valid, generalisable knowledge about interventions, going beyond the immediacy of programme evaluation. Programme research therefore is appropriate in its own way to the study of social programmes. For many, the purpose of an evaluation is "to improve the programme" (formative evaluation). But a programme can also be improved in the long term, through an improved understanding of the nature of the social problem it addresses, how it is implemented, how the programme processes are supposed to bring about change, etc. The paper now turns to how programme research contributes to this.

Programme effects

Programme outcomes ought to be a concern in evaluation. Taylor (1995), for example, asserted the need "to establish whether or not a particular project works and, if so, to disentangle the various factors which contribute to making it effective" (p. 3). In this

regard it is useful to distinguish between immediate or short term effects (which we could call outcomes), and long term or ultimate effects (call them impacts). In designing research to determine outcome or impact, confusing the two might lead to prematurely positive or negative findings of the programme's effectiveness. For example: an evaluation of a teacher training programme may find that it resulted in more group-based activities in the classroom, and that positive changes in the attitudes of pupils occurred. This might be what the programme intends to accomplish, but the evaluators and programme staff should be aware that these would be intermediate or proximal effects: in the long run, the programme is aimed at improving pupil learning. Whether the programme accomplishes this would remain an open question at this point of the evaluation.

Causality and research design

When we speak of outcomes, intermediate or ultimate, we imply cause-and-effect relationships, and we need to design studies which will be able to detect such relationships with a certain amount of confidence. In research design terms, randomized experimental designs, and quasi-experimental designs, exist to facilitate causal inference (Cook and Campbell, 1979). Although it is impossible to treat the philosophical and methodological concerns involving causality in this brief paper, one comment has to be made: in experiments and quasi-experiments we are dealing with causal inferences, and not matters of fact. This kind of research constructs an argument, using logic, methodology and empirical data, that the intervention caused the observed changes. This is achieved via comparisons, across time or across persons: in the latter, for example, the treatment or programme (the independent variable) is given to one group of individuals, and witheld from another. The performance of the two groups is then compared in terms of the variable of concern - the dependent variable (see Lipsey,1993).

Experiments and quasi-experiments are designed to promote internal validity, the validity of our inference that the relationship established between the variables of interest is a causal one. These designs attempt to rule out alternative interpretations of the observed effects, by exercising control over aspects of the study: keeping out extraneous forces; determining who receives a particular treatment at a particular time; and eliminating threats to internal validity (see Cook & Campbell, 1979). To put it differently: experimental and quasi-experimental designs are powerful methods of controlling for variance. That is, they attempt to maximise some sources of variance, and minimise others. Internal validity then is about minimising sources of unwanted variation.

The threats to internal validity, and how to counteract them via research design, are familiar to most - Cook and Campbell's (1979) book contains a long list. In the rest of the paper, I would like to focus on three groups of factors to consider in designing an outcome study, relevant to the educational context of the JET conference. These relate to the respondents (subjects) in the study, the treatment itself, and measurement.

Subject heterogeneity

The number of people/ data points taken from the population is acknowledged as a major consideration in the design of a study: larger samples, randomly selected, represent the population better. But another factor to take into account is how widely individuals differ on a variable of interest in a population. Pupils in Grade 4 are different from pupils in Grade 8; teachers in rural schools differ from teachers in urban schools. Educational programmes acknowledge this: they seldom are designed for all school children, but rather for children in a specific grade; or for teachers with a specific educational profile.

The heterogeneity of subjects on the dependent variable introduces a problem for the researcher, since it makes it more difficult to detect a difference between those who have been exposed to the programme and those who have not. Heterogeneity in terms of variables like age, school grades, gender, socio-economic status, etc. spreads the dependent variable scores out more, and makes the experiment less sensitive. Fortunately, there are sampling techniques available which are of assistance. For example, the sample could be stratified into

different levels of age, grades, etc. Within each stratum the subjects then are more alike, but the study still includes a range of subjects. This increases the sensitivity of the study to detect an effect, and addresses the extent to which the results could be generalised.

Treatment

Variability in treatment and control conditions in experimental and quasi-experimental designs introduces experimental error. When the experimental procedure itself produces unwanted variablity, the validity of the study is compromised - because we are looking for variability as a result of the treatment.

In programme evaluation, Yeaton and Sechrest (1981) have referred to this as the integrity of treatment. That is, the degree to which the programme is delivered in a uniform manner to the appropriate recipients at the right time. Programmes are complex entities, containing many elements. As a result, as Lipsey (1993) points out, delivering the programme in a standardised, uniform manner to all recipients is often not possible in the field. The best safeguard then is to actually measure or determine the treatment received by subjects in the various conditions. Taylor (1995) implicitly supports this position, by arguing that quantitative outcome studies need to include a description of what happened in the classroom, to fill the gap between the programme-as-intended and the improved knowledge of pupils. One needs to know how the programme was delivered in reality, to be able to link its outcomes with "what happened in the classroom".

A further consideration from Yeaton and Sechrest's (1981) paper is the strength of the treatment. It stands to reason that stronger treatments should produce larger effects. The difficulty of course is what is a strong treatment in delivering a curriculum? It would appear that in-service teacher training programmes are based on the assumption that they increase the strength o f curriculum by increasing the expertise or skills of the teachers. This is a reasonable assumption, but we can extend our consideration of strength of treatment to ask about the strength of teacher training itself as an intervention for a specific educational situation. Given the problems identified in the schools targeted by the NGOs as reviewed by Taylor (1995), one might ask: Is in-service teacher training a strong treatment for those problems?

Measurement

Any measuring instrument must have validity - it must measure the characteristic of interest and not something else. Lipsey (1993) has added that it must also have validity for change, it must be sensitive to the changes induced by treatment. One kind of measurement typically is not sensitive to change: psychometric measures of things like aptitude, personality, and IQ, are developed to assess the extent to which each individual has a certain characteristic. They are therefore inappropriate for testing whether pupils have mastered something. Edumetric measures are designed primarily to detect change; e.g. a test of mathematics. It is this kind of measure which typically will give an educational programme maximum opportunity to show its effectiveness.

A second standard requirement of measuring instruments is that they must be reliable; that they must not show random error in repeated measurement of the same quantity. Unreliable measures have more variability, and variability of this kind reduces the ability of the experiment to detect differences. Taylor (1995) identified four categories of data used in the evaluation studies he reviewed: self-report data, via questionnaires and interviews; project documents; classroom observations; and instruments for assessing cognitive skills. The reliability of these instruments is often compromised by well-known potential weaknesses. Questionnaires are influenced by fluctuations in attention, motivation, comprehension, etc.; observers doing behavioural observations in a classroom may see different things, or use different standards; record keeping systems may lack quality control, or may be incomplete or inconsistent in what they report. In all of these examples the measuring instrument introduces unwanted variability, which reduces the chances of finding significant effects.
One of the most straightforward ways to overcome these measurement difficulties is to avoid mono-method bias. Each construct or

variable must be measured in more than just one way. If the variable of interest is, for example, aggressive behaviour in class, it is possible to ask teachers to rate children in terms of aggressiveness, to conduct observations in the classroom, and to administer paper and pencil tests attempting to measure the same construct.

Programme theory

Many readers by now would have said: That is all fine and true, but in the real world of educational programme evaluation the treatments are complex, the human beings to whom they are delivered are even more complex, and it is impossible to control field conditions. Thus it is almost always very difficult to attain the levels of control required of experiments.

It is in this regard that the emphasis of evaluation research on programme theory becomes important. In the so-called black box evaluations, the evaluator is concerned simply with the relationship between input and output, and not with how the input led to the outcome. Programme theory is concerned with the causal chain of what happened in the black box; with specifying how the programme worked. Some change processes must have taken place which link the treatment of the programme with the desired outcomes. These transformational relations between treatment and outcomes, and the contextual factors within which they occur, are of interest to programme theory.

The argument here is that a causal theory of how the programme works is of practical importance, but in the longer term. In a nutshell, this is an argument for why academic research (or programme research), is of use in programme evaluation. If we understand and are able to specify how a particular intervention works, we are in a much stronger position to improve the programme itself, as well as to improve the design of an outcome study. Programme theory gives guidance to the evaluator in a number of ways. For example: it contributes to specifying relevant populations for the treatment and samples for the evaluation; it assists in identifying variables likely to be important, and ways in which to measure them; it distinguishes between intermediate and ultimate effects; it focuses attention on treatment strength and integrity; it allows stakeholders to understand why and how a treatment will or will not work; it provides insight into the social problems targeted, and the appropriateness of the programme; and much more.

Conclusion

Programme research, and its value in a case like teacher training, applies both to the "front" end and the "back" end of the evaluation process.

At the front end, it takes us back to a set of important questions about the nature of the educational situations for which the teachers were being trained. In addition, it re-directs attention to what the skills are that the teachers bring to the training in the first place. What then are the skills they need to refine/ develop from the training? How does one move them from skill point A to skill point B? In the long run, knowledge of this sort speaks to how to design effective programmes, select appropriate trainees, etc., and not simply directly to evaluation.

At the "back" end of the evaluation, the question is how to go beyond piecemeal results from individual, project-specific evaluations to some broader knowledge base about what we know generally about a type of programme/ intervention. This was one of the major considerations for JET calling this conference on programme evaluation. There is more than one way to address this question, but in terms of the argument developed in this paper, it is easy to point out how "theory" can act as a vehicle for accumulating knowledge from individual evaluations. This means attempting to build some model of the change process represented in an effective programme, including the input factors, the outcome effectiveness issues, and contextual factors. This demonstrates how a theory-based approach can improve programme evaluation as well as programme design. As such, it does much more than address the immediate concerns of programme stakeholders. However, project personnel, programme sponsors, and society, are well served by good, knowledge-based programme design.

In conclusion, the main points of the paper can be summarized as follows:

• programmes have effects, and these should be taken seriously;

• research designs are available to enable inferences about causal :relationships between treatment and outcomes;

• these designs provide internal validity control, increasing confidence in the inferences made;

• programme theory strengthens research design; and

• theoretically informed and well designed studies contribute to our knowledge of programmes, and improve programmes in the long run.

REFERENCES

Cook, TD and Campbell, DT (1979) Quasi-experiments. Design and analysis issues for field settings. Chicago: Rand McNally.

Cordray, DS & Lipsey, M (1986) Program evaluation and program research. Evaluation Studies Review Annual, 11, 17-44.

Lipsey, M (1993) Theory as method: Small theories of treatments. New Directions for Program Evaluation, 57,5-38.

Taylor, N (1995) INSET, NGOs and evaluation: A Review. Paper delivered to the Kenton-at-Settlers Conference, 27-30 October.

Yeaton, WH and Sechrest, L (1981) Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. Journal of Consulting and Clinical Psychology, 49,156-167.

WALKING THE LINE: JUDGEMENTS OF VERIDICALITY AND FAIRNESS IN TEACHER EDUCATION EVALUATIONS

Professor Johan Muller, University of Cape Town

"The judge must render a just sentence: the historian must provide a truthful account."

<div align="right">Davidson, 1995: 307</div>

"Politicians and intellectuals are as useful as they are dangerous. More precisely: they are dangerous as a result of the same autonomy that conditions their utility. Nothing special. But surely 'something different'".

<div align="right">Pels, 1995: 98</div>

There are two major classes of judgements that are brought to bear in programme evaluations. The first is the class of veridical judgements, judgements of truth or veracity. The second is the class of judgements that encompasses justice and fairness. Every evaluation will have components of each class of judgement, although their 'mix' will differ from one evaluation to another. [1]

Veridical judgements are based on demonstrable evidence. The rules for gathering the evidence as well as the rules of inference that string the evidence into an informative narrative, are the rules of the scholarly or epistemic community. We commonly call this cluster of judgements 'methodology', although that is an oddly unsubstantial word for the cluster of competencies that is the toolkit of the evaluator. Let us say that the duly inducted and duly licensed adept of veridical judgements is the scholar. The scholar's accountability, is first and foremost to her scholarly community and to their rules. Her licence depends on a trustworthy exercise of those rules, and anyone hiring the scholar must be able to trust that these rules will be adequately applied. Johann Louw's paper in this collection has taken us in exemplary fashion into the heart of the world of validity of the scholar. It is a terrifyingly rigorous world but it is that rigour which permits our confidence in the entailments of the evidence.

The second is the class of judgements that encompasses justice and fairness. The adept of this class of judgements is the adjudicator. If the scholar must provide a truthful judgement, the adjudicator must provide a just, fair, moral and politically acceptable judgement. One of the big differences between the adjudicator's and the scholar's judgement is that, whereas the scholar looks primarily to her scientific community, the adjudicator is accountable to a range of constituencies and stakeholders, upon whose judgement of fair treatment the success of an evaluation rests. Every evaluation is thus composed of a set of veridical judgements which assesses, in terms of the evaluation brief, what is the case: and a set of judicious judgements that establishes what is fair and just. The truth game traffics in knowledge; the justice game traffics in power, or the distribution and allocation of privilege (accountability is the recognition of apt privilege). The former leans towards the academy; the latter leans towards politics.

We may say, then, that one measure of the adequacy of any evaluation is the degree to which the two regulative ideals (truth and justice), and the two classes of judgement, are represented, combined or connected in the evaluation. If we accept that, then the big question is how?

---

[1] A good deal of the argument in this paper was thrashed out with Johann Mouton at Lamberts Bay.

It will not often be the case, even though it is not in principle impossible, that the competencies of the scholar and those of the adjudicator are adequately represented in one person. Why might that be? First, there are different cognitive interests at stake here, the one pursuing truth (what its the case?) the other pursuing justice (is it fair?). Mastering the skill of good judgement takes time, training and practice, and it will be difficult to change hats or discourses in a perfectly accomplished way. Also, whereas the scholar faces the epistemic community only, the adjudicator faces the stakeholders, whose interests and conceptions will frequently differ [2]. Moving smoothly from the epistemic community to the stakeholders and back again, repeatedly, will take some doing. It is not impossible, but it will The rare. Of course, the exceptions prove the rule. Scholar politicians like Jan Smuts and Lenin come to mind here. Most often, of course, the evaluation contracting agency has either a truth or a justice-centred view of evaluation anyway, and then it is far easier to locate a seemingly appropriate person that will match that often tacit view. I will return to the problem with such strategies in a moment.

At the forerunner to this conference (JET Workshop, Sept. 1995) two evaluations were presented. Both were good. evaluations, but they were good in different ways. The first produced an exceedingly impressive set of evidence which showed quite clearly what the project being evaluated had been doing and with what effect. The second produced highly plausible, but intuitive, judgements that went well beyond the evidence, yet seemed right, and were illuminating. We didn't use these terms then, but in retrospect we could certainly say that the first evaluator leaned towards scholarship, the second relatively towards justice.

Let us agree for the moment then that it is desirable to have some or other mix of good judgements in an evaluation. How might an evaluation contractor go about ensuring this? It seems to me that there are four options logically available:

1. The first one would be to select a demonstrably competent scholar to provide evidence, and leave the judicious judgements to the funders or their Board. There are three difficulties here: first, it leaves rather too much initial shaping power to the scholar; more importantly, funder Boards may have impressive constituency representatives on them, but this does not necessarily ensure judiciousness. judiciousness, like scholarliness, is a hard-earned skill. Of course, the Board members could undergo some special training (though they often think their representativeness accords them all the political acumen they need). But there is a further difficulty: they, at the end of the day, represent one stakeholder, while the judicious judgement requires a balancing of accounts across multiple stakeholders. And judiciousness is not something that can be delivered by a vote. By and large, this strategy will not stand a good chance of delivering justice on a regular and reliable basis.

2. A second strategy would be to put a scholar and an adjudicator together on a team. The partnership would be forged from the start. Getting the two to cooperate might not always be easy, but this model is a definite feasibility. What might preclude it is the practicalities of the matter. Agencies often find it easier to contract an evaluator.

The pragmatic evaluation contractor might well set out to try to optimise and to choose an evaluator with acknowledged skills in both, to at least a minimum level of

---

2 This point is an ideal-typical, not an historical one. In point of fact, higher education institutions globally are becoming more porous and less insulated from the outside world. To that degree, they are having to account for and justify the expenditure of public and private funds, as well as their teaching and research priorities, in ways they never have had to before. And this is probably a benign development (see Scott, 1995). It remains, however, a matter of relative emphasis. The boundary between the university and the world is not infinitely porous. There comes a point aft er which the knowledge game may become so worldly, so preoccupied with the ideal of justice, so driven by worldly accountability, that the ideal of truth ceases to have regulatory force. At that point, the truth game, indeed the university in its social mission, loses its singular role and becomes indistinguishable from other worldly or practical knowledge games of pragmatic interest-brokerage (see Stehr, 1992).

acceptability. In practice, however, one of two possibilities is probable:

3.   a scholar with more (or less) of a 'feel' for adjudication; and conversely,

4.   an adjudicator with more (or less) knowledge of research and methodology. The former is rather more likely where the scholar has decided to make a career out of evaluation and has resolved to take seriously adjudication as a process of judgement. However, where this is more dream than reality, a regress to option one is not unlikely. The latter, likewise, is more of a hope than a reality in many cases. My guess is that a political person who has learnt the skill of dealing with shifting plural constituencies is unlikely to spend the time required to serve the single and sometimes unforgiving protocols of the epistemic community.

These then are the logical possibilities. What are some of the implications to be drawn? First, let's look at what the evaluation contractor might conclude from the above.

•   The major lesson is that the identification and selection of evaluators is a business that goes far beyond simply finding out who is available. It has to do with identifying the range of judgements most pertinent to the case at hand. Is methodological sophistication needed? Political delicacy? Whatever the answer, no evaluation will ever be without some measure of both.

•   The second lesson is that briefs to evaluators must be far more carefully drawn up. Exactly what evidence is required? Exactly what stakeholders are key to the success of the project? It is undoubtedly so that a major reason for the slippage between evidence and opinion that characterises so many programme evaluations in South Africa, and that makes them, in retrospect, so singularly uninformative, is the fact that the briefs tended to blur the requirements. • A third lesson is also implied here: the evaluators must be part of finalising the briefs. Why is that? The contractor has usually identified with either the stakeholders, or more rarely, with the scholarly community. It is all too easy to give the other domain of judgement short shrift in the brief. For example, the scholar should be part of stipulating exactly what evidence would constitute sufficient grounds for judgement in the case at hand.

What might scholars or potential scholars who want to do evaluations deduce from this story?

•   The first, and I would say overriding, requirement for a programme evaluator is a sound, credible, methodological competence. This is the most saleable asset. It is what the contractor will really rely on (even though there are contractors out there who do not yet realise this).
•   If a career is envisaged, rather than just doing it on the side, as is often presently the case, then developing political sensitivity and judiciousness will be a good option. My guess is that this will not happen overnight. And it is certainly not in the interests of scholars to espouse a judicious competence they do not have.

Projects are stakeholders too. They have a legitimate interest in developing the criteria which will be required in the brief and operationalised in the assessment. Up till now, any input by the projects has been in a bilateral fashion. The time is ripe for projects to get together and develop a collective position on the way they are to be judged. My sense is that funders will welcome this. However, all of this really only makes sense if funders are committed to developing long term relationships with projects. Short term funding in a crucial sense undercuts the development of stable evaluation environments.

If the two forms of judgement are not explicitly recognised and built into the evaluation from the start, there are three equally unsatisfactory outcomes that could occur:

1. Either because of the inclination of the contractor/ funder, or the inclination of the evaluator, or both, the evaluation can be implicitly or explicitly designed as an exercise in scholarly judgement only. Here, veridical judgements and evidence are all that are deemed necessary. We might say that this is evaluation as expertocracy.

2. On the other hand, the funder may have (in fact many funders have) designed evaluations as exercises in political legitimation. Here, evidence and veridical judgement is deemed unnecessary, and is downplayed or even set aside. The stakeholder reputation or street credentials of the evaluator is what, in the main, is contracted here. We may say that this is evaluation as pure politics, for in the absence of warrantable evidence, it is hardly an evaluation at all.

3. Finally, there are some evaluators who want to blur the distinction between veridical and judicious judgements, who won't want to make the distinction at all. The result can only be a goulash of judgements whose warrants are impossibly confused. It behoves the contractor to examine very carefully what the skill base of these evaluators is. At the end of the day, the worth of the evaluation lies in the demonstrable accountability, which is always dual:
   - to the evidence and rules of methodology,
   - to the stakeholders and the criteria of fairness and justice they can live with. The daunting task of the contractor is to broker a harmonious marriage of those two in the case of every evaluation.

REFERENCES & SELECT BIOLOGRAPHY

AHRS (1995) Social Knowledge for Societal Transformation. HSRC.

Davidson AI (1991) 'Carlo Ginzburg and the renewal of historiography'. In J Chandler, Al Davidson & H Harootunian (eds), Questions of Evidence, Chicago: University of Chicago Press.

Ginzburg, C (1991)'Checking the evidence: The judge and the historian'. In J Chandler, Al Davidson & H Harootunian (eds), Questions of Evidence, Chicago: University of Chicago Press.

Fels, D (1995) 'Knowledge politics and anti-politics: Toward a critical appraisal of Bourdieu's concept of intellectual autonomy'. Theory and Society, 24, 79-104.

Popkewitz, TS (1996)'The denial of change in the process of change: Systems of ideas and the construction of national evaluations'. Mimeo.

Scott, P (1995) The Meanings of Mass Higher Education. Buckingham: The Society for Research into Higher Education.

Stehr, N (1992) Practical Knowledge. London: Sage.

DOES TEACHER DEVELOPMENT WORK? TRUE CONFESSIONS OF A HARDENED EVALUATOR[1]

Professor Jonathan D. Jansen, University of Durban Westville

Teacher development delivered through NGOs (non-governmental organisations) in the form of in-service education (INSET) does not produce significant learning gains in the classroom. One evaluation report after another has delivered the same finding: that while INSET provides important motivational benefits to practising teachers, and begins to influence the behaviours of participating teachers, such programmes simply do not translate into learning gains for students. Yet millions of rands have been invested in teacher development programmes of South African NGOs since the mid-1970s, increasing dramatically through the 1980s, and levelling-off in the first half of the 1990s. Why has NGO-driven teacher development not delivered?

This paper will argue that the underlying reason for this state of affairs can be traced to the doorstep of both teacher development programmes and evaluation studies. In short, teacher development programmes have serious design flaws; on the other hand, evaluation studies have operated as if these design flaws did not exist. What is recommended is both a reconceptualisation of teacher development and a corresponding re-design of programme evaluation. Within this understanding, I address the broader conference question: how can the methodologies used in evaluations be refined to ensure that evaluations contribute to the body of research in this field?

I should depart briefly from the main storyline with some biographical information which sets the context for the main arguments pursued in this paper. In the past six years I have been involved in various ways in the evaluation of educational programmes delivered by NGOs with the specified goal of improving the development of in-service teachers in science, mathematics and language education (mainly English). In this paper I will summarise the results of at least twelve different evaluation studies in which I have either conducted the evaluation, or supervised the evaluation with little direct involvement in the data collection process, or acted as supervisor and been active in the associated fieldwork for the evaluation.

One should begin with an understanding of the historical context within which evaluation emerged in South Africa. Evaluation of teacher development programmes (and indeed of most development work in general) emerged initially as a simple bureaucratic requirement of international and, to a lesser extent, national funding agencies. Evaluators in these early times asked simple questions about programme delivery (how many training workshops were conducted?) and participation rates (how many people were reached?). Evaluation was typically conducted by a small team of consultants who, in the case of international funding agencies, were mainly from the country represented by that agency [2]. Such evaluations examined results using a crude model of programme outputs, leading to the delivery of an evaluation report in a three-week period on average. In the main, evaluation was something done by an external agency to a

---

[1] 1 should thank my colleague, Renuka Vithal, for inspired comments on an earlier draft of this paper. She is absolved, in true Catholic tradition, from any responsibility for the pain this paper might cause.

[2] In a survey of 177 evaluations done in 50 countries, Snyder & Doan (1995) find a similar tendency to hire foreigners in the case of the evaluation of American foreign aid.

NGO with the focus on simple numerical indices of performance. Unsurprisingly, NGO staff working in teacher development typically responded with fear and disdain for the alleged judgemental nature of evaluation, the lack of participation and 'process' in the evaluation, and the inevitable threats to programme continuance given the constraints under which such organisations operated during the anti-apartheid years. This simple bureaucratic interest in evaluation ended quickly at the start of the decade. In what could be called "the post-Boesak era of NGOs," the question of impact was raised as a serious and central concern in aid agency evaluation work during the 1990s. With what consequences?

There are at least three major findings from the evaluation of programme impact across a range of education NGOs concerned with teacher development. I will briefly list and discuss each of these findings, with a more detailed set of reflections on the findings concerned with the impact of teacher development programmes on student learning. I will then return to my central thesis: that the absence of impact is a fault of both development work and evaluation strategy in the post-Boesak era.

1. NGO-driven teacher development programmes impact positively on the motivation of teachers across school subjects and grade levels.

In NGO INSET programmes for science, languages and mathematics, teacher participation and involvement is high. Participants consistently record appreciation of
a) time available for teacher development;
b) the hands-on, inquiry-oriented approaches to teaching;
c) the focus on student-centred learning;
d) the making available of "supplementary materials" which often constitute the only available resources for teaching and learning;
e) the broadening of the repertoire of familiar teaching methodologies and
f) the opportunity to work in a collegial environment, learning from fellow teachers working in similar struggles and contexts.

If NGO INSET initiatives articulated teacher motivation as the exclusive goal of these interventions, then the impact of such programmes has been optimal. But few self-respecting NGOs would make such a claim. There are other, deeper goals.

2. NGO-driven teacher development programmes show only moderate impact on the instructional behaviours of participating teachers.

There is some evidence that teachers teach differently as a result of their participation in teacher development programmes. In science and mathematics classrooms participant teachers have invariably moved away from teacher-dominated instruction. Small group learning is evident; students handle ("manipulate" according to the manuals and syllabus) materials; teachers pose questions rather than simply rattle off statements of fact; and some part of the instructional time per scheduled 'period' involves self-directed student learning. There are, however, four important limitations of the available evidence.

First, the available data is often limited to self-reports by teachers or principals who participated in these INSET programmes. This is problematic. When asked whether, for example, they use student-centred teaching strategies, teachers invariably claim that they do. When the same teachers are observed in the classroom, it has been found that they do precisely the opposite i.e., classes are organised around the teacher and her teaching. At a first glance, it appears that teachers are untruthful. But this is not the case. The question "do you use student-centred teaching strategies" appears to be heard as a question about what is done under ideal conditions. Teachers in their responses therefore project their ideal teaching situation, not what they actually do in a real classroom. The point of this discussion is that self-reporting on intended impact on teaching behaviours is grossly misleading with respect to what actually happens.

Second, when evidence is gathered through classroom observation, much of this data on teachers' instructional behaviours has been gauged through scheduled or anticipated evaluator visits focused on the particular interventions of a designated NGO, often with

a staff member of the relevant NGO present? This means that there is a reasonable chance that teachers may be teaching to the observers on a well-rehearsed topic. In several classroom visits connected with INSET evaluations, we have observed that the fluidity of the lesson in terms of teacher action and student responses has been so smooth that it could only have happened through prior rehearsal. However, some evaluators have argued that, even if contrived, the available data does suggest that teachers can change their instructional behaviours in directions suggested by NGO INSET programmes.

Third, most of the available data on teachers' instructional behaviours has been collected through one-off evaluator visits rather than on the basis of sustained observations over time. Little can be said in existing evaluations about the long-term impact of INSET on teachers and teaching. Evaluation designs are not geared to collecting longitudinal data nor are they based on pre-existing baseline data; accordingly, the evidence on whether changes in teaching, where observed, are actually sustained, is not knowable through existing information.

Fourth, when evaluations include observational data, observations are often not cross-checked with data gained through intensive interviewing or the analysis of written work to probe for the understandings which may or may not underpin an observable behaviour. There is some devastating evidence that this failure of evaluation design may have led to some erroneous assumptions about the nature and degree of impact on teaching behaviours. Consider the following examples observed in two separate primary science classrooms where the teachers had been exposed to in-service training. In the first classroom (A) the in-service training prepares teachers for involving students in doing acid-base tests. The evaluator enters the classroom and sees small group learning; hands-on inquiry with students using test-tubes, indicators, and a variety of "unknown solutions"; the teacher acting as facilitator of the learning process, moves around from group to group, raising critical questions, and generally leaving the students to follow the outlined procedures. At a first glance (that is, with the focus on behaviours), this is John Dewey or Paulo Freire teaching at its best. Until the evaluator looks more closely ... that is, with the focus on understanding. The students are indeed in the required learning mode. But they are filling each of the test-tubes with the indicator (of which an unbelievably large amount is available) and using the dropper to run-in singular droplets of the unknown solution. Unfortunately, many evaluations have retained a focus on behaviours at the expense of understanding and this constitutes the single most important threat to impact evaluation studies in South Africa.

In the second classroom (B), the teacher is 'implementing' a component of an in-service programme concerned with introducing students to practical work in learning about electricity. The following dialogue ensues around a static electricity experiment:

Teacher: do you see the paper jump to the ruler?
Students:yes (in chorus)
Teacher: why did the ruler jump?
Students:[no response]
Teacher: the ruler jumped?
Students:[no response]
Teacher: [smiles and moves to the next point]

The teacher in classroom B clearly posed the right question. But she could not pose the follow-up questions which would indicate both competence in posing good probing questions based on unexpected responses and

3 The NGO staff member presence constitutes a double-edged sword. On the one hand the NGO staff member may influence what is observed by dictating teacher expectations; on the other hand, the NGO staff member provides vital access to isolated schools and often complex school environments hostile to outsiders; essential translation services are also sometimes provided by such a staff person.

students' understanding of what they saw. Evaluators observing this teacher might record teacher-initiated questions as taking place and the experimentation with materials on the subject of static electricity But the evaluators would not be able to make any judgement about the depth of understanding of the subject matter on the part of either the teacher or students.

In short, as a result of poor evaluation designs and the ensuing lack of good evaluation data, there is little evidence that INSET changes and sustains teaching practices. On the other hand, available data does show some impact on overt teaching behaviours at a particular time and in a particular context (Stevens 1995, for example).

3. NGO-driven teacher development: programmes show little to no impact on learning gains among students.

The unspoken assumption in INSET is that the direct training benefits to teachers translate into learning gains for their students. Students are the ultimate beneficiaries of INSET. They are expected to learn more and better as a result of their teachers' exposure to INSET services. Is this assumption tenable? The answer is emphatically negative. Why?

INSET in South Africa is designed around trainers and teachers. Accordingly, where impact studies have been commissioned, they have focused on immediate training benefits to the teachers. Motivational questions and behaviour change questions feature prominently in South African INSET evaluations. Student learning is an afterthought in the design of INSET programmes and the accompanying evaluation, although, as stated earlier, the linkage to learning is assumed.

The few studies that have explored student: learning gains have only demonstrated motivational benefits. Students 'manipulate' the learning materials; students work enthusiastically in groups; students are engaged with the new teaching strategy introduced; students respond well to questions etc. What is not clear is what: happens inside the heads of students as they learn science or mathematics; is there a qualitative difference in what is learnt as a result of the intervention? This deeper level of impact is not explored. In one unique pilot study examining the cognitive effects of learning in a multimedia science programme, the evaluators signalled some difference in the quality of learning in the new programme (Behr & Perrold 1993). A few other studies have been able to demonstrate concrete learning gains through INSET.[4] Even in these cases the data is suspect. The methodology adopted often did not involve intensive interviewing of students, in-depth observations or longitudinal designs which could have provided such evidence. In short, there is an absence of 'depth data' to support the claims about learning gains in the classroom.

The evidence suggests, rather, that since teachers seldom have developed depth understanding of what is offered in centre-based, one-off or sporadic workshops, it is unlikely that students have benefitted in terms of learning gains (see earlier examples)[5] Furthermore, the recitation and routinisation of innovative tasks among students suggest that while new rules have been learnt, new understandings have not necessarily developed. In addition, despite the inquiry-based approach of all INSET programmes, in few classrooms observed was there much evidence of student-initiated questions, especially questions related to the substance of the lesson. There was certainly no evidence of students challenging teacher explanations. But there is some evidence of students posing management questions e.g., how long should the assignment be?

[4] The outstanding study designed by David Bateson (1995) on comparative learning gains in the Science Education Programme (SEP) is one such exception. But SEP is also one of the few programmes which combines centre-based training with intensive classroom-level support.

[5] The design of INSET programmes is not irrational. There are heavy demands on a NGO to go for breadth (extend basic training to more teachers) rather than depth (provide more intensive training for a small group of teachers) in service delivery. Similarly, the design of evaluations is also seldom irrational. There are heavy cost implications in doing multiple observations or the shadowing of teachers over time to provide the 'depth data' alluded to.

In sum, evaluation designs do not allow for good data to be captured relevant to student learning. And where such data does exist, it suggests that students may be learning new ideas in old ways; adapting an innovative approach into a mechanical task; collapsing inquiry into routine.

It should be clear from the discussion thus far that this state of affairs can be located squarely in relation to the twin factors identified at the start of this paper: the failure of INSET and the failure of evaluation designs.

What does this mean for the generalisability of knowledge generated through INSET evaluations and its contribution to the larger body of knowledge on evaluation and research? This is a dangerous question. First, because it assumes evaluation should say something to a broader community of social scientists and educational researchers. This is problematic. Evaluation is done for a specific audience i.e., an internal audience (e.g., Boards of Trustees) or an external audience (e.g., donor agency). The design of the evaluation therefore parallels the interests of that audience. For example, USAID may require an impact study focused on learning gains within a six-week consultancy. Designs are compromised to the point of dishonesty to meet the deadline. The maxim becomes, 'what you see is what you get.' That is, within the six-week period the US-based consultants take four weeks to figure out how a country could possibly have 19 education departments. Such is a very different audience to an academic group where elegance, parsimony, creativity and the production of cutting-edge information are much more important in research design than in the case of evaluations where designs are often bent out of shape to meet ridiculous deadlines, donor expectations or, for that matter, to satisfy grantee pressures.

Second, generalising information generated in evaluations is problematic because it assumes evaluation can say something to a broader academic audience. Evaluations are typically done in specific contexts, in specific cultures, in specific constraints. The knowledge generated within these specific locales must be treated with extreme caution.

If generalisation is a priority, such a goal must only be pursued if:

1.   there is deep and transparent contextualisation of the evaluation issues under discussion so that a reader can understand not only what is being claimed but where it is being said and under what conditions;
2.   there is some minimum set of parallel conditions in the development context e.g., an evaluation being conducted in rural Brazil where teachers are teaching large classes in under-resourced schools could inform evaluators doing similar studies in rural Northern Province under similar conditions;
3.   there is some degree of uniformity in the evaluation designs being pursued in the two different evaluation contexts, so that the knowledge claims associated with one design are not spuriously related to claims made within a very different evaluation design; and
4.   there is some approximation of INSET delivery strategy across the two evaluation contexts.

To illustrate this last point: you simply cannot compare the data on learning gains from the Danish Life Science Project in Namibia (which involves extensive centre-based training in the life sciences followed by intensive classroom-based support to participating teachers over a period of at least one year) to similar science INSET training programmes delivered over the odd week-end to whomever shows up. In other words, it is almost impossible to seek generalisations about learning gains given the many layers of complexity and difficulty in the way evaluations are conducted in South Africa and the generic problems associated with comparative evaluation designs.[6]

The hard-nosed quantitative evaluator (of which there are thankfully few) would propose a simple solution. Design an experimental or quasi-experimental study, assign subjects to an experimental or control group, and the ambiguity of findings on learning gains is resolved. This kind of design

---

[6] In a purist sense, the only evaluations worth comparing are those producing 'effect sizes' based on similar statistical procedures; this process is called meta-analysis.

has both ethical (how to assign students on the basis of non-treatment conditions) and methodological problems (how to attribute gains recorded given a multi-faceted INSET programme). Moreover, such studies merely extend evaluator blindness on the complex and unpredictable school and classroom level processes which shape the sought-after outcomes (such as learning gains). This is critical information to an evaluator concerned with programme improvement, the overriding rationale for conducting programme evaluation in South Africa. But even summative evaluation would be incomplete without an understanding of the contribution of the different components of a programme and their interactive effects at the classroom level.

This is a problem which is also poorly understood in the large-scale production function studies characterising much of the research on schools in developing countries. Achievement (or what we have called learning gains) is explained in terms of discrete 'inputs' (such as textbooks or teacher salaries or laboratories) without a clue about the kind of textbooks, the qualities of teaching, the contingencies of classroom contexts or the transactions between teachers and learners which influence achievement (see extended critique in Jansen 1995b). Some have responded by calling for an input variable called "processes" (Lockheed & Komenan 1989). Others have called for multi-level modelling which disaggregates effects at, for example, school, classroom and individual pupil levels (Riddell 1996) as a way of clarifying 'what causes what.' But both these responses miss the important point that quantifying process in a real-school environment: is a contradiction in terms. That is, the critique is assumed to be targeting a procedural dilemma rather than an epistemological problem.

I have painted a dismal picture about the actual contribution of INSET to learning and the possible contribution of evaluation studies to the body of knowledge in the field. I propose the following ways out of these dilemmas.

1.   That funders in their evaluation designs, in the short term, recognise and accept the internal goals of an INSET programme as valid. Such goals may primarily be to broaden the repertoires of teaching strategies available to underqualified science teachers rather than a definitive improvement in learning gains among students of those trainees.

2.   That INSET programmes, in the long term, be re-designed to focus on learning gains as a primary goal. This means rethinking simplistic modes of delivery to meet demand in favour of more sophisticated programmes seeking depth (more and better training over sustained periods of time) rather than breadth (more of the same, spread thinly over time and space).

3.   That comparative evaluation studies be designed and conducted with the deliberate goal of examining the value of rival programmes not with the goal of discarding the less successful (although this might be necessary in an extreme case) but with the aim of transferring lessons learned across projects. This means that comparison is built into the design (and the selection of comparable cases) from the start rather than in retrospective studies on set designs.

4.   That evaluation studies be designed to capture 'depth data' on learning through a new ensemble of data collection procedures which include clinical interviewing, sustained observations, longitudinal designs and the routine collection of baseline data against which to measure changes over time. We have exhausted the use, and can no longer justify the utility of, mindless questionnaires, 'lucky dip' observations (essentially, dipping into a classroom without any sampling plan or evaluation focus so that what you see is dependent on the event of the moment) and idealised self-reports. But this requires that those commissioning evaluations set different conditions for the conduct of evaluation than those producing the 'quick and dirty' variety.

5.   That evaluation studies set a new standard for evaluation reports e.g., produce the richly contextualised narratives which bring to light powerful findings on impact

beyond statistical summaries. But this requires setting new standards of validation such as articulated by Yvonna Lincoln and Egon Guba (1991; and see Shadish et al., 1995) for qualitative studies in evaluation.

This paper has declared a curse on both the house of the funding office (such as USAID, JET etc) and the INSET NGOs (especially those involved in teacher development). Both houses have legitimate concerns; both have serious flaws in their operating assumptions about programme evaluation and in-service teacher training. Changing strategy and discourse about both these issues will bring about a more informed debate on understanding how (and whether) things change as a result of our development efforts.

SELECTED REFERENCES

Atkin, Myron (1995) Evaluation Report, Handspring/ PSP. IEQ: Durban.

Bateson, David (1995) Evaluation of the Science Education Project. Johannesburg: SEP

Bhika, Sharma (1996) Evaluation of a Selected Study Guide for Distance Education INSET. Masters Dissertation, University of Durban Westville.

Dlamini, Thulisile (1996) Evaluation of the Mathematics Centre for Primary Teachers. Masters Dissertation, University of Natal (Durban).

Jansen, JD (1995a) Classroom Observations in the Eastern Cape, Handspring/PSP. IEQ: Durban.

Jansen, JD (1995b) 'Effective Schools?' Comparative Education, 31(2): 181-200.

Lincoln, YS (1991)'The Arts and Sciences of Program Evaluation.' Evaluation Practice, 12:1-7.

Lockheed, ME & Komenan A (1989) 'Teaching Quality and Student Achievement in Africa: the case of Nigeria and Swaziland.' Teaching and Teacher Education 5: 93-113.

Ntshingila-Khoza, Roseline (1995) Teaching in South Africa. Observed Pedagogical Practices and Teachers' Own Meanings. IEQ: Durban.

Perrold, Helene et al. (1995) Evaluation of the Handspring Trust for Puppetry in Education. Johannesburg/ Durban: Joint Evaluation by IEQ, PSP, HTPE and Helene Perrold & Associates.

Riddell, Abby (1996) School Effectiveness Studies. Paper presented to the International Conference on Teaching and Learning, University of Bristol, 6-9 January.

Shadish WR et al. (1995) Guiding Principles for Evaluators. California: Jossey-Bass, Summer no. 66.

Snyder, Monteze & Doan, Peter (1995) 'Who Participates in the Evaluation of International Development Aid?' Evaluation Practice, 16(2): 141-152.

Stevens, Floraline (1995) Preliminary Impact Evaluation Report of NGO Educare Training in South Africa. IEQ: Durban.

SCIENCE UNDER THE SPOTLIGHT

Professor Peter Weingart, Bielefeld University

1. Changing Legitimation of Science

The notion of science into which the present generation of scholars was institutionalized entails a number of characteristics which may undergo a fundamental change in the next few decades to come.

These pertain both to the ways in which society looks upon science as a social enterprise which produces privileged, certified knowledge, and the ways in which this enterprise functions internally. The former relates to forms of legitimation of knowledge production, the latter to the kind of knowledge that is produced. Both are, as will become apparent in the subsequent argument, inherently related.

Science has, ever since its emergence as a social institution, served political (and utilitarian) ends whenever feasible and called upon. Yet, we now witness a new trend in science policy: a call for a closer orientation to national goals, and linked to this the challenge to increase the efficiency of research and teaching institutions to open up to evaluation from outside in order to be more accountable to the public, i.e. the taxpayer or stakeholders as the funders of the enterprise.

Pressure to align more closely the expenditure:; for research with more immediate national needs affects a basic understanding of the process of research and development which has served as legitimation of government R&D budgets for almost half a century. This understanding dates back to Vannevar Bush's model according to which science could be differentiated into basic research and applied research, and that both activities could be neatly kept apart. The further assumption was that basic research was oriented to the discovery of natural laws, it could not be planned since by definition it was directed to secrets yet to be unraveled. Between basic research, applied research and the ultimate transformation into concrete technologies (development) would lie a time span which could in some cases be shortened but under no circumstances be eliminated. In other words: the categories of research activity represented a continuum which was at the same time a unilinear sequence.

The implication of this model for funding was that basic research was considered crucial for the national innovation system, and yet was elusive to any strategic planning. It was to be funded on its own merits with the expectation of the ultimate social utility of a broad knowledge base thus created. This arrangement, established immediately after the war under the impression of the contribution of science to the war effort and the dangers of its prolonged military organisation, responded to the unique situation in the US. Yet, given the American leadership in science it provided an ideological legitimation for the support of 'pure research' which owed much of its continued adherence to the cold war even in other parts of the world. With the cold war having come to an end this legitimation seems to be waning. A more immediate proof of the utility of research is demanded. The political pressure to provide this proof is mounting across all Western industrialized countries. For some time now the categories on which Bush's model rested have not seemed appropriate, especially their sequentiality, as experience shows a relationship o£ many reverse connections, feedback loops and the like.

Transdisciplinary and interdisciplinary research fields with social and/or economic interests as their foci emerge and disappear. Thus, a familiar organisational scheme is given up, for good reason. 'Pure research' thereby loses its self-determined character, even if no similarly operative scheme is in place.

There is another reason why the existing contract between the science establishment and society is being reviewed. Bush's model called for considerable trust on the part of the public to be invested in science. The public mostly had only the scientists' word that what they were doing was ultimately for the common welfare. However, parallel to a fundamental democratisation of society in which the privileged role of an estate (DK Price) is becoming increasingly fragile, the hitherto accepted authority of the science establishment came to be questioned on account of a growing awareness of risks and ethical dilemmas. The new demands on science are diverse, and it is by no means clear how they are going to be operationalized. But among them is the call for external evaluation which is already being implemented in a number of countries. It is worth looking first at the significance that the new evaluation schemes have for the understanding of science and its funding.

2. From internal to external evaluation

The most fundamental characteristic of science is its independence from outside direction. In a sense, this is self-evident: science is the ultimate source of objectified, formal knowledge. Thus, there can be no higher plane from which to judge it. Interference from outside into the manufacture of "truth" is of little use. This is true regardless of the many failings that can readily be observed in the doings of scientists. Science is self-referential in the sense that scientists have to relate to the research of other scientists for their own research. To refer to the authority of God or that of the government would be regarded as inappropriate to advancing the argument. Likewise, only scientists themselves determine what is science, not the ministry of education. The self-referentiality of science finds its institutional expression in the mechanisms of internal control, i.e. peer review. This mechanism which has been in place since the inception of modern science operates in scientific journals, in the organisation of conferences, in the distribution of funds through national or private foundations. It is the control of experts by experts who alone have the same expertise and are therefore in a position to exert that control. It assures, in spite of occasional malfunctions, the self-critical direction of research and thus the 'quality' of knowledge production. But 'peer review' also has one fundamental problem which only becomes apparent when the funds that science asks for grow scarcer and the demand for accounting for their use becomes stronger. Under such conditions the suspicion that peer review is self-serving inevitably appears, and it is enhanced by both reports on failures of the system, be it fraud, misjudgements, and disclosures to the public of the sometimes very small circles of peers that are evaluating each other. The peer review system which exemplifies the trust relationship between science and the public thus becomes strained.

For at least a decade the demand for outside evaluation of scientific research has been mounting. This demand is primarily directed at the universities as the largest, most important, and most costly institutions of research and teaching. Given their enormous expansion in the 1970s the question now is how to make them more efficient and cost-conscious. Several countries, led by the Netherlands, the UK, and recently Switzerland, have implemented evaluation schemes with that objective. It must be recognised that the introduction of such schemes is a major change in the conception of science. On the other hand this development should be seen in perspective: not only science but also the entire public sector, i.e. government administrations and community services are now subject to the demand for proof of efficiency and productivity. One has to remember that to this day the legitimation of governments' achievements is based on expenditures of public funds, not on an evaluation of the results achieved with them. Evaluation of science, therefore, is not part of an anti-science movement.

Any attempt to evaluate the achievements of research from outside, i.e. by laymen, would still be fraught with the problem that it does not have command over the requisite expertise, and would therefore not be acceptable to the researchers. Externalization of evaluation could not mean, therefore, introducing evaluative criteria different from those inherent to science. In this respect basic science differs from utilitarian production of knowledge. Just as 'peer review' applies only to basic research, i.e. the production of fundamental knowledge, the whole issue of internal vs. external evaluation only arises in that context. This leaves aside for the moment the issue of the borderline riot always being unequivocally determinable between the types of research.

The problem was and continues to be to find ways by which the evaluation would be based on criteria which are inherent to research and would not compromise the 'internally' established expertise, and yet could be applied from the 'outside'. This is an overriding principle of the evaluation of research and the basis for the search for 'science indicators'. The principle instigates that the products of research be transformed so that they can be measured. In that way research is made accountable to the public, and the sources of mistrust stemming from the obscurity of peer evaluation are avoided.

The solution to the problem as it is particular to science is the use of literature data banks as sources of indicator construction. Like any indicator construction (e.g. in social analysis or economics) the underlying rationale is that the measurable data must represent the product which is to be evaluated. Any indicator, thus, presupposes a 'theory' about the relationship between the object to be measured and the representative measure. In the case of science such indicators are the number of publications, the number of citations, honors and prizes received, invitations to conferences and combinations of all of these. It is no accident that the discussion over the use of these indicators has focused on the problem of validity: do these indicators really measure quality of research or put more generally, do they measure what they are supposed to measure? In the evaluation community this is the standard question, and it is approached by testing as well as by advancing the theoretical debate on the 'internal' operation of research. In this discussion finer differentiations have been worked out: the number of publications as such reflects 'activity' rather than quality per se; citations are best understood as a measure of 'recognition'; absolute numbers have to be normalised for different disciplines because citing patterns differ substantially from one field to the other; time frames are important because the 'half life' of a publication is very different in the natural sciences than in the humanities, and so on.

Beyond these refinements, however, the chief justification for the use of these indicators in science is a dual one. Firstly, they reflect the products of research including their recognition by peer review. In this sense they are products of the internal process of knowledge production and creditation. Secondly, they lend themselves to quantification and can, thus, be made visible to the outside without experts having to intervene. Both of these reasons have greatly enhanced the legitimacy of indicator-based evaluations of basic research.

The latter should not be taken to mean that experts have no role in this process: the function of quantitative indicators in evaluation is not to replace expert judgement but to provide an objectified basis for that judgment, a background to which it can respond, an orienting framework which leaves room for interpretation but constrains it. Typically, quantitative indicators are used in conjunction with expert judgement, and, in effect, they enhance the latter's legitimacy, as well. If expert judgement is passed on the basis of quantified data it rules out to a great extent manipulation, for it makes interpretation of the data accessible. Also, it helps the experts to see things which otherwise they may not be aware of, patterns which they cannot possibly see from their limited subjective perspectives. No-one can have confident knowledge about all the literature in one's field. In fact, it has been proved in validation exercises that results of

quantitative indicator-based evaluations vary as little or as much as qualitative judgements, and that the latter are improved in scope by the former.

To underscore this point a concrete example should be given of an arrangement encompassing the use of quantitative indicators and of expert judgement in tandem. The Swiss Science Council has commissioned several bibliometric studies of the state of different research fields such as physics and clinical medicine. These studies provided the Council with an overview of the publication activity of all relevant institutions in the country (i.e. universities and independent as well as industrial institutes working in the field). Such an analysis provides a 'map' depicting the centres of activity (and those most recognised in terms of citations received) as well as the more marginal institutions. Subsequently an international evaluation committee of experts was established which undertook an in-depth analysis of the field. They had the results of the bibliometric analysis in hand and could follow up conspicuous data. In the end the committee prepared its report with suggestions as to where to increase and where to withdraw funds or which infrastructural measures should be taken in order to improve the situation.

Note that in this somewhat simplified arrangement the quantitative indicators are not used exclusively to decide over the life or death of institutions but rather in conjunction with expert judgement. It guides this judgement but the experts may also arrive at a different evaluation when their detailed analysis reveals particular reasons for the situation as described in the quantitative analysis. At the same time it forces both evaluators and those evaluated to confront the picture as presented by the bibliometric analysis. Thus, the process of expert evaluation is depoliticised to a large degree and rendered more efficient.

3. Evaluation and complexity

Any evaluation of a social activity, whatever it may be, boils down to judgement which, in essence, is a reduction of complexity. As mentioned above this reduction takes the form of indicators when quantification is aspired to. Quantification has the great advantage of being generally communicable and therefore open to general criticism. This does not mean that numbers are unequivocal and objective per se. In the case of indicators the interpretative input is apparent in the 'theoretical' input which is their basis. It is because this 'theory' is not always available or made explicit in the process that one hears almost invariably the argument from opponents to evaluation by indicators that the quantitative measure cannot grasp the 'essence' of the phenomenon under scrutiny. While it may be true that 'essence' is never measurable, proponents of this argument are never able to define what it is and thereby are subject to allegations that they evade evaluation altogether. The 'quantitative/qualitative' dichotomy raised in debates over evaluation is an ideological one, and it is fruitless. The only fruitful issue between opponents and proponents of evaluation is whether the theories of representation are deemed plausible or not.

It is also obvious that the more complex the phenomenon to be evaluated, the more difficult it is to establish reliable theories on which to base indicator construction. A case in point is the evaluation of programmes of the enhancement of research and/or teaching, of the advancement of innovation, or more specifically of the development of particular technologies. Programme evaluation is usually directed to a whole set of policy measures implemented by funding agencies and/or governments. It is designed to measure the effect of such programmes in order to decide whether expenditures of funds be continued or not. Given the complexity of the relationship between the intervention and the system in which it is to have an effect there is much confusion about goals, desired effects, ways to record them and measures to evaluate them. A recent survey of programme evaluations in Germany prepared for the Ministry of Science and Technology revealed that so far no programme evaluation actually conformed to the standards usually required for evaluation. The reasons for this deficiency are varied: recency of

introduction of evaluations, lack of data, scarcity of funds, inexplicitness of goals. This finding is corroborated by the review of evaluations carried out by Taylor in 1995. While the complexity of programmes of social intervention of any kind is recognised the design of evaluation schemes may, in fact, help to make them more efficient. A few principles can be stated which are fairly consensual among evaluation scholars and social scientists:

- Beyond the ideologically tainted debate over quantitative and qualitative measures it must be clear that judgements which are not at least intersubjectively communicable are obviously worthless. If participants and administrators in a programme cannot agree with their funders or outside evaluators on the programme's achievements, the measures used are useless, and they will ultimately lead to a loss of credibility.
- The issue is not whether evaluation measures (indicators) should be quantitative or qualitative but that they be adequate to the subject matter that they are intended to represent. The more tenuous and fragile the 'theory' on which they are based the more important it may be to support them with qualitative interpretation of data. Specific attention must be paid to the collection of data, i.e. information about the effects achieved by the programme. They depend, of course, on the definition of the programme's goal(s).
- It is absolutely crucial that the goals of the programme are clearly stated. Only with a precise definition o£ goals is evaluation of effects possible. Vague goals are, in effect, a carte blanche to spend funds on activities defined by the party receiving the funds, and the funder should not expect accountable results. Of course, it is possible and quite customary to place the responsibility to do so in the hands of the recipient. In those cases evaluation in the strict sense is not possible.
- A common defensive argument is that while implicit or professed goals may not Qd~._ be reached the programme may have latent achievements not initially stated or even known. (Process value as against product value). This may well be the case but must be measurable as well if it is not to serve as a purely evasive argument.
- A further condition of evaluation is comparability. Any evaluation must have a baseline on which to judge impact, effect or achievement. This condition is often the most difficult to achieve both for lack o£ data and funds or time. In many cases the creation of a control group is either impossible or ethically questionable.
- Self-evaluation is a source of data at best, but is inevitably self-serving. It should be used to collect information. It should not be used as a measure of success.
- The main objective of evaluation is the improvement of the activity under scrutiny, not the execution of control for its own sake. Thus, evaluation, if properly understood, is a mechanism enabling the organisation to be evaluated to learn. Organisational (and individual) learning is only possible if feedback is provided. Therefore, two principles are essential in the application of evaluation procedures:
    a) the results of any evaluation have to be fed back to those evaluated, and
    b) evaluation and decisions deduced from the results of the evaluation have to be separated.

To summarize: There should be a defensible relationship between the objectives of funders, programme objectives, and evaluation measures. Much of this linkage will have to be negotiated during the design of programmes. Experts executing programmes have knowledge which the fenders are lacking. This puts trust on them. But in the last analysis the experts will have to demonstrate their achievements in terms accessible to stakeholders. If they cannot do so that trust will be withdrawn to the detriment of future efforts of good will.

REFERENCES

N Taylor, (1995) INSET, NGOs and Eva luation: A Review. Paper delivered to the Kenton-at-Settlers Conference, 27-30 October.

For further details of indicator construction and application in the evaluation of research:

P Weingart, M Winterhager, Die Vermessung der Forschung, Frankfurt, 1984

P Weingart, R Sehringer, M Winterhager, Which Reality Do We Measure? in: Scientometrics Vol. 19, 5-6, 1990, 481-894;

CASE STUDY

IMPACT EVALUATION OF THE INDEPENDENT TRAINING AND EDUCATIONAL CENTRE

Submitted to the Joint Education Trust

and the United States Agency for International Development

November 1995

TEAM MEMBERS

Jennifer Bisgard - Team Leader
Ms Bisgard is the Director - Educational Services of Khulisa Management Services. Based in Cape Town, she co-ordinates the educational services of Khulisa, with particular focus on impact evaluations. She has a masters degree in Development, Social Change, and Economics from Johns Hopkins University School of Advanced International Studies. Prior to joining Khulisa, Ms Bisgard was the Supervisor - Basic Education with the United States Agency for International Development (USAID) in Pretoria where she co-ordinated $30 million of education assistance.

Irene Mpolweni - Evaluator Western Cape
Ms Mpolweni is employed by the Grassroots Educare Trust based in Cape Town. She is the co-ordinator of the Field Services Department which includes planning, budgeting and evaluating as well as managing a team of field workers. Prior to this post, Ms Mpolweni was involved in Organisational Development Training for Grassroots. Ms Mpolweni has a Social Science Degree from the University of Cape Town where she majored in Sociology and Personnel Management. She is currently pursuing a diploma in Adult Basic Education, also from UCT. Ms Mpolweni is originally from the Eastern Cape and speaks Xhosa fluently.

Michael Ogawa - Computer Specialist
Mr Ogawa is the Education and Management Information Systems Specialist, based in the Johannesburg office of Khulisa Management Services. Prior to joining Khulisa, Mr Ogawa administered and implemented the Canadian Fund for Local Initiatives (Canada Fund), the United Nations High Commissioner for Refugees (UNHCR) Education Assistance Programme, and the World University Service o£ Canada (WUSC) Refugee Scholarship Programme in Swaziland. Mr Ogawa has a BSc in Mathematical Computing from the University of British Columbia and has partially completed an MSc in Statistics.

ACRONYMS

ABE	Adult Basic Education
ANC	African National Congress
CASME	Centre for the Advancement of Science and Mathematics Education
CIE	Catholic Institute of Education
CRIC	Career Resource and Information Centre
DET	Department of Education and Training
ECD	Early Childhood Development
EDO	Education Development Officer
EOF	Equal Opportunity Foundation
EU	European Union
EXPI	Experimental Group 1
EXP2	Experimental Group 2
EXP3	Experimental Group 3
HRD	Human Resource Development
IDT	Independent Development Trust
IEB	Independent Examinations Board
IEQ	Improving Educational Quality (a USAID project)
IHRD	Institute for Human Resource Development
INSET	In-service Teacher Training
ITEC	Independent Training and Educational Centre (formerly Independent Teacher Enrichment Centre)
ITRC	Independent Teachers Resource Centre
JET	Joint Education Trust
JP	Junior Primary
JPTD	Junior Primary Teaching Diploma
JTC	Job Training Centre
LMAP	Language Methods and Programmes
LP	Lower Primary
MBSA	Mercedes Benz of South Africa
MIS	Management Information System
MRU	Mobile Resource Unit
MTD	Management Training Department
NGO	Non-Governmental Organisation
OLSET	Open Learning Systems Educational Trust
PETF	Provincial Education and Training Forum
PRESET	Pre-Service Teacher Training
PROTEC	Project Technology
PTA	Parent Teacher Association
PTC	Primary Teaching Certificate
PTD	Primary Teaching Diploma
QSP	Quality Schools Project
RDP	Reconstruction and Development Programme
READ	Read, Educate and Develop
SANCO	South African National Civic Organisation
SEP	Science Education Project Std Standard Sub-ASubStandard A
Sub-B	SubStandard B
TOPS	Teacher Opportunity Programmes
TSP	Thousand Schools Project
USAID	United States Agency for International Development
Wits	University of the Witwatersrand

PREFACE

This evaluation of the Independent Training and Educational Centre (ITEC) was jointly funded by the joint Education Trust (JET) and the United States Agency for International Development (USAID). The study examined the impact of two key ITEC programmes: Lower Primary Project and School Management and Governance Training. In addition, the overall impact of ITEC on the community and the extent to which the centre could serve as a replicable model were examined.

To maximise efforts, Khulisa Management Services worked co-operatively with the USAID-funded Improving Education Quality (IEQ) project which has been working with ITEC to design evaluation instruments as well as gathering data to assess impact. For the Lower Primary Project, the Khulisa team co-ordinated the evaluation process utilising questionnaires and observation continua developed by IEQ and ITEC. For the School Management and Governance Training portion of the assignment, ITEC gathered data on the participants while Khulisa spot checked the data collection during the field work. To examine the overall community impact, ITEC's staff collected data using a form that was jointly designed with Khulisa. The Khulisa evaluation team was responsible for examining the extent to which ITEC can serve as a replicable model, by conducting interviews.

Khulisa Management Services would like to acknowledge the co-operative effort of ITEC in contributing to the completion of the evaluation assignment. ITEC provided a considerable level of effort on the part of its the staff in the evaluation process. Its staff was consistently open and helpful in assisting the evaluation team. Khulisa also acknowledges the co-operation of the IEQ staff who designed the Lower Primary data collection instruments, trained the facilitators in collecting data, supervised data collection and selection of schools, and, finally, was very supportive throughout the analytical process. The opinions expressed in this report are those of the evaluation team and not of the donor agencies. Similarly, errors in this report are the responsibility of the authors alone and should not be attributed to the donors, individuals and/or institutions interviewed for this evaluation.

1. OVERVIEW

The Independent Training and Educational Centre (ITEC) is a large non-governmental organisation (NGO) located in the Eastern Cape serving the educational and training needs of disadvantaged South Africans. Funding was provided by the United States Agency for International Development (USAID) and the Joint Education Trust (JET) to commission Khulisa Management Services to conduct this impact evaluation. The purposes of the ITEC evaluation are to:

a) Examine the impact of the Lower Primary Programme;
b) Provide a review of the Management and Governance Programme;
c) Assess the overall impact of ITEC on the community; and
d) Provide an estimation of the replicability of ITEC in terms of management, sustainability, costs, and networking.

Only the evaluation of the lower primary programme is included here. A full version of the evaluation can be obtained from JET or ITEC.

2. BACKGROUND

2.1. ITEC HISTORY

During the educational crisis of 1985, grassroots political protest prevented teachers in the Eastern Cape from teaching. As a result, a community effort to upgrade teachers during this "down time" commenced. The teachers expressed a strong need for an educational, teaching and resource centre. Several community members formed an association, conducted a needs assessment, obtained a building and commenced fund-raising activities. The culmination of this community action was the launch of the Independent Teacher Enrichment Centre (ITEC) in 1987. Since then, ITEC has grown from its original staff of two to 40 full-time staff and another 40 part-time staff (mostly tutors). Originally focused only on the needs of schools, teachers and pupils, the organisation over time shifted emphasis to embrace life-long access to education by all members of the community. In 1993, the name was changed to the Independent Training and Educational Centre to reflect this altered focus.

ITEC's August 1995 strategic planning exercise outlines the following target populations:
• Education Gatekeepers: Teachers, principals, INSET/ PRESET trainers, education development officers (EDOs - formerly inspectors), subject advisors, and government department officials;
• Community: Leaders, families /parents/ guardians, and school governing councils; and
• Youth: In-school and out-of-school youth.

The ultimate objective of reaching these target populations is to support the children of the community.

Over time, ITEC has run a variety of programmes to promote teacher competency and student attainment in a well-managed environment with access to resources. The current programmes address needs such as school secretarial services' (typing and printing examinations), principals' forums (to encourage sharing of skills and resources between schools), governance training, teacher training, and programmes that focus specifically on the needs of students (PROTEC, Science Fair, Saturday schools, and the Mobile Resource Unit). Community programmes include the literacy programme, the Job Training Centre and the resource centre. The result is an educational and community centre where access to programmes and skills is enhanced.

Although ITEC's headquarters are located in East London, management has chosen to focus primarily on rural schools that are particularly disadvantaged. A notable achievement has been its research into conditions in farm schools in the Border region conducted in conjunction with the Border Early Learning Centre (now called the Community Child Development Centre). This research highlighted the need for training of lower primary school teachers who are often faced with very limited resources, lack of formal PRESET training (many teachers only have attained Std 8 plus a two year teaching diploma) and multi-level classes.

[1] At one point in its early years, ITEC responded to the needs of the schools for well-groomed environments by providing mowing and garden services for a marginal fee. This achieved one of ITEC's aims, namely to increase the level of pride and community involvement. Once this had been achieved, the programme was discontinued with schools maintaining their own gardens.

2.2. CONTEXT

The Eastern Cape Province is one of the most impoverished in South Africa, The Eastern Cape government, based in Bisho, faces a very difficult task of integrating two former homelands, the Ciskei and Transkei, along with four other departments of education. The combined system comprises 6,000 schools.

The 1994 statistics gathered by the Provincial Education and Training Forum indicated:

- Provincial population of 6,090,504;
- 2,100,000 children in school of which 1,754,881 are in primary school;
- 57,000 teachers;
- 6,000 schools (3,257 primary schools) and 2,500 early childhood centres/pre-schools; and
- 23,000 department officials (those working in the Department of Education and Culture, but not teachers or principals); and
- 20 Subject Advisors for the primary sector.

The government services available in the Eastern Cape Province are generally inadequate to support the needs of teachers. Currently, there is one subject advisor for every 1,590 primary teachers, and many of these advisors focus exclusively on senior primary support. The low number of subject advisors supporting primary education indicates a lack of technical support for junior primary school teachers. The large number of department officials within the system is the result of the amalgamation o£ the Ciskei and Transkei and the interim constitutional guarantee of jobs until 1999. Many of the 23,000 bureaucrats are clerks who perform duplicate duties and lack the basic facilities and resources which would make them more effective. The government has four INSET centres for the province which are equipped to run courses, but do not have the staff or infrastructure to provide school-based INSET. Eighty percent of the province's population lives in areas classified rural. The Eastern Cape road system is not well-developed or maintained. As a result, it can take more than nine hours to drive from one end of the province to the other.

Teachers travel great distances to reach schools located in remote areas that usually lack electricity or access to telecommunications. In one case, a teacher reported travelling 15 hours a week between home and school. It is not uncommon for pupils to walk up to 10 kilometres to reach school, or for parents to enrol nine or ten-year-old children in sub-A as the long distances to school are difficult for younger children.

During the period of the evaluation fieldwork, a serious scandal of the Eastern Cape School Nutrition Presidential Project had just been exposed. Funds for the school nutrition scheme had been misused and misappropriated, resulting in the suspension of the school nutrition programme. When the programme was abrogated, attendance in rural schools dropped dramatically with some schools reporting absenteeism of 40 to 50 percent according to newspaper reports.

This context and the articulated needs of communities are the main motivators for ITEC to develop programmes such as the Lower Primary Programme described below.

3. LOWER PRIMARY PROGRAMME

The Lower Primary (LP) programme is ITEC's largest initiative, employing approximately 40 percent of ITEC's staff. The programme is managed by the LP Project Leader who has two deputies in charge of training and materials development respectively. The training team consists of seven facilitators who provide workshops and follow-up support for teachers. The materials development team includes two curriculum developers, one material developer, one marketing assistant, and one materials/ printing assistant.

3.1. OVERVIEW OF THE PROGRAMME

The LP programme provides non-accredited courses for teachers of Grade 1 through Std 1 who desire to improve their classroom competence. Currently, 265 teachers from urban, peri-urban, and rural areas (including farm schools) in the Greater East London and Ciskei regions are taking part in the programme. In 1995, this number was augmented by a further 500 teachers comprising 408 Quality Schools Project (QSP) teachers, 32 teachers in conjunction with the University of the Witwatersrand (Wits) Rural Project and 60 teachers in association with

[2]"Meeting the Educational Expectations of our Democratic Society: Proposal for a Provincial Teacher Education/INSET Process" by the Provincial Education and Training Forum, Eastern Cape Province, November 1, 1994, page 4.

the Catholic Institute of Education (CIE) in Kokstad. In the latter two cases, ITEC provides the workshops and materials while Wits and CIE provide administration and follow-up support. The programme is conducted over three to four years during which eight modules are completed. Two-day workshops are held during the first two or three weeks of each school term (30 percent of the programme). Field support by facilitators follows during the next six weeks of the term (the remainder of the programme). Only field support takes place during the last term. Facilitators provide classroom support to each teacher in the programme at least three times a year. Workshops focus on learner-centred methodologies, and on individual and classroom organisational and management skills. The support aspect includes monitoring progress, demonstration lessons and the effective and optimum use of teaching aids/materials. An optional two-day workshop on how to make teaching aids (including appropriate scripting) is held during the winter vacation in July. During the fourth term, teachers attend a needs/ expectations forum to prepare for the following year. Community members (governing councils and PTAs) are also invited to participate in the forum.

3.1.1. Modules
The key modules (teaching competencies) of the training programme are:
- Module 1: Creating a child-centred learning environment;
- Module 2: Implementing the school readiness programme;
- Module 3: Good handwriting skills incorporating designing worksheets and workcards for group work;
- Module 4: Implementing investigative mathematical principles;
- Module 5: Learning and teaching language (methodology);
- Module 6:]integrating language learning across the curriculum;
- Module 7: Identifying learning and health problems, and community health awareness; and
- Module 8: Assessment without writing examinations.

Each module is supported by teaching notes suggesting new ideas, and a set of basic teaching aids which support and complement the interim junior primary curriculum. Many of these materials are available in all 11 official languages; all materials are available in Xhosa, English and Afrikaans (the main languages of the Eastern Cape).

3.1.2. Nature, Length, Time and Frequency of Intervention
Each year, three modules, each comprising 10 hours over a two-day period (30 hours per annum), are held in a centre convenient to teachers. In 1995, a total of 15 venues were utilised. In a number of cases, workshops were staggered so that half the teachers in a school could attend the workshop while the others teach, and vice-versa the following week. ITEC has received verbal permission from the Department of Education in Bisho to hold workshops during school hours.

Winter school consists of one optional module over two days (centre-based), constituting an additional 10 hours.

During the pilot (research and development) phase from 1991 through 1993, there were no monitoring visits conducted during the first two years, and only one monitoring visit was held in 1993. In 1994, the programme changed to its present modular approach. Field-based monitoring visits are held three times per year for each teacher. This constitutes 4.5 hours of classroom support per teacher per annum. There were only two monitoring visits in 1994 due to unrest prior to the April 1994 elections. There were two monitoring visits conducted in 1995 prior to this evaluation (August 1995).

The programme requires between three and four years to complete. The speed at which the teachers complete the programme depends on their ability to master requisite competencies specified in the modules and to complete the self-evaluation portion of the programme. A self-evaluation instrument forms part of every module.

Most of the teachers have completed one year of the training programme in its present form. There are a number of teachers who have been part of the ITEC LP pilot since 1991, but have not participated in the revised modules. Table 1 summarises the number of teachers who have completed each module, and the number of hours of support received.

3.1.3. Goals of the Programme

The goals of the LP Programme are to improve the quality of education and learning in junior primary schools through:

• Providing teacher training and support by a team of trained facilitators who spend 30 percent of their time facilitating workshops and 70 percent providing field support; and

• Designing, testing and producing support materials and classroom resources.

3.2. EVALUATION METHODOLOGY

This evaluation followed a three-stage process to measure attitudinal change, behavioural change and impact on children's learning. Table 2 provides a summary of the evaluation methodology used by the evaluation team to measure changes and impact. The evaluation is partially based on the assumption that if teachers' attitudes change and if the teachers are provided resources, then their behaviour will change, resulting in better, more effective pupil learning.

3.2.1. Data Collection

The Khulisa evaluation fieldwork benefited from the extensive training and assistance provided to ITEC by the USAID-sponsored IEQ project. IEQ's brief is to build capacity of USAID basic education grantees to monitor and strengthen the impact of their programmes. The IEQ worked with ITEC to develop the impact assessment instruments to identify LP's impact:

a. On teachers:
   • Attitude changes;
   • Methodologies becoming learner-centred; and
   • Use of resources and support materials,
b. On pupils:
   • Utilisation of materials and resources;
   • Interaction with each other to enhance learning; and
   • Meaningful interaction with the teachers.
c. On classroom:
   • Materials appropriately displayed;
   • Being conducive to child-centred learning, and
   • Seating arrangement focused on child-centred learning.

Instruments used to measure the foregoing included:
   • Teacher interview protocols administered to 11 ITEC teachers;
   • Classroom observation protocols examining 40 ITEC classrooms and 10 control classrooms; and
   • Teacher background questionnaires administered to 40 ITEC teachers and 10 control teachers.

To establish pupil learning gains, ITEC developed assessment activities to measure mathematics and language competencies following the syllabi of sub-A, sub-B and Std 1. All children were Xhosa first language speakers.

Because some of the classrooms where pupil assessments were administered were located in farm schools having multi-level classrooms and as one ITEC teacher had shifted from teaching sub-A to Std 2, some Std 2 pupils were given the Std 1 assessment activity. Presumably, this would have biased the Std 1 assessment results upwards. For this reason, and the fact that there were no Std 2 classrooms in the control group, the Std 2 pupil results were excluded from the statistical analysis of the Std 1 assessment activities. To ensure fairness between control and ITEC classrooms, the ITEC facilitator (rather than the class teacher) provided instructions to the children and monitored the pupil assessment activity while the

Table 2: Evaluation Methodology

	First stage	Second stage	Third stage
NGO teacher development programme	Attitudinal change	Behavioural change	Impact on children's learning
Evaluation instruments	Questionnaire interview	Classroom observation	Assess pupil performance

[3] The figures in the table reflect the highest (not cumulative) level of achievement

Table1: Summary of LP Modules and Teacher Support [3]

INTERVENTION	NUMBER OF TEACHERS WHO HAVE COMPLETED EACH STAGE AS OF AUGUST 1995
MODULE 1	341
MODULE 2	932
MODULE 3	194
MODULE 4	213
MODULE 5	294
MODULE 6	107
MODULE 7	0
MODULE 8	0
3 HOURS SUPPORT	163
6 HOURS SUPPORT	78
7.5 HOURS SUPPORT	24

teacher was asked to complete the teacher questionnaire in the staff room. This procedure eliminated a potential bias that could have affected the pupils' results.

Finally, as an additional control, the evaluation fieldworker accompanied each facilitator to collect data in seven ITEC classrooms and three control classrooms. This methodology provided independent verification of the data collection process as well as important observations.

3.2.2. Selection of Classrooms

The evaluation team visited 40 experimental group classrooms' and 10 control group classrooms. The classrooms were randomly selected in consultation with ITEC to ensure a fair demographic representation of ITEC's coverage in the Eastern Cape. Classes were selected to represent a range of environments with experimental categories established to represent the number of years of ITEC training as indicated below: The control classrooms have had no input from ITEC. In one case, a control classroom at Mzamowethu was eliminated from the study as the school was a recipient of other programmes from ITEC including the Mobile Resource Unit, READ and OLSET (in 1994). Chat classroom was replaced by one at Sinomonde School.

The classrooms were first divided into three categories (rural, farm and urban), as shown in Table 4, and then sub-divided by standard levels.

While official policy states that Sub-A pupils are to be admitted to school at the age of six years, the average age for Sub-A pupils is 7.32 years and 8.64 years for the experimental and control groups respectively. The fact that older children are in Sub-A could indicate that parents are enrolling their children at a later age due to economic reasons or due to the fact that the distances to walk to school are too much for younger children. This could also indicate that there is a high drop-out and repetition rate at this grade level.

[4] Forty experimental schools were visited by the evaluation team; one set of experimental school data from "school 13" is missing and, thus, data from 39 experimental schools was analysed.
[5] The data from "school 13" from EXP1 is missing; thus, data from only 12 EXPI schools was analysed.

Table 3: ITEC Enrolment Figures

Year Started	# Registered	Continued in 1995	Dropped Out of Programme
1991-93 (Pilot Group)	197	24 *	173
1994	230	78 **	152 ***
1995	163	163	0
Total	590	265	325

* Experimental Group 1 (EXP1) consists of 13 [5] classrooms involved from the pilot stage to the present and was selected from this group of 24 teachers.

** Experimental Group 2 (EXP2) consists of 14 classrooms involved in the programme since 1994 and was selected from this group of 78 teachers.

*** Experimental Group 3 (EXP3) consists of 13 classrooms involved in ITEC workshops in 1994 only and then dropped out of the programme. EXP3 was selected from the 152 teachers who participated in 1994 and then dropped out of the programme (68 of the 152 teachers dropped out after the first module and were not selected for EXP3).

Table 6: Profile of Experimental and Control Group Pupils

	Experimental	Control
# of Pupils	1,182 (48.14% female; 51.27% male)*	306 (48.04% female; 51.60% male)**
EXP1	389	N/A
EXP2	437	N/A
EXP3	437	N/A

* Gender was not specified for seven experimental group pupils.
** Gender was not specified for one control group pupil.

Table 4: Classroom Selection

	Experimental	Control
# of Classrooms Where Assessment Activities Administered	39	10
Urban Classrooms	12	4
Farm Classrooms	8	1
Rural Public Classrooms	19	5

3.3. RESULTS
3.3.1. Profiles of Experimental and Control Groups

Table 5: Pupils Assessed During the Evaluation

	Experiment	Average Age	Control	Average Age	Desirable Age
Sub-A Pupils	394	7.32	131	8.64	6
Sub-B Pupils	293	8.97	51	8.53	7
Std 1 Pupils	495	9.81	124	8.95	8
Total	1182	8.85	306	8.74	n/a

3.3.2. Changes in Classroom Practice

3.3.2.1. Teachers

3.3.2.1.1. Attitudinal Changes

The teacher questionnaire asked teachers to agree or disagree on a scale ranging from one (strongly disagree) to five (strongly agree) if they had realised improvement in their teaching abilities during the last one to three years in the following areas:
• Teaching skills and ideas;
• Confidence in teaching ability;
• Motivation to teach; and
• Knowledge of subject.

A fifth statement queried teachers if they had not experienced any changes in their teaching abilities. A composite score was then calculated for each teacher by averaging their responses to the five statements. Experimental group teachers consistently reported that they strongly agreed that ITEC training had helped them improve in the foregoing areas. Although control group teachers also reported improvements in those areas, ITEC-trained teachers were associated with a significantly higher composite score of 4.14 compared with 3.43 for the control group teachers ($p<0.005$) e. In other words, ITEC-trained teachers reported a higher level of improvement in these areas than the control group teachers.

Some of the teachers interviewed indicated that the ITEC training made them realise that there was a need to change their teaching habits and to learn and utilise new teaching methodologies. Other teachers mentioned that the ITEC training assisted them with their Sub-A teaching, as they had never before received any specialised Junior Primary training. One teacher said that the training had "made me care more about the child's interests". Another said that "by having students involve themselves in independent studies, I have time to assist slow learners".

The majority of the experimental group teachers interviewed agreed that the ITEC training helped improve their self-confidence and communication skills. They also indicated that they were now able to give advice to and help other teachers, suggesting a spin-off benefit for other teachers who are not involved in the programme.

3.3.2.1.2. Teaching Methodologies/Behavioural Changes

The teacher questionnaire asked teachers to rate how often they used the following methods on a scale from one (rarely) to three (most of the time):
• Hands-on activities;
• Pupil-centred teaching;
• Group work activities;
• Independent work by pupils; and
• Active participation by teachers.

Experimental group teachers reported that since receiving INSET programme training, they used these methods most of the time (composite score = 2.77). Control group teachers also indicated they used these methods most of the time (composite score = 2.62). This difference, however, was not statistically significant. In other words, ITEC-trained teachers did not use these methods significantly more than the control group teachers. The classroom observation measured the teachers' use of a variety of teaching methods on a scale ranging from one (teacher uses one method that does not involve learners) to four (teacher uses more than two teaching methods which all involve learners) - see the graph below and Table 7. The experimental group used a significantly greater variety of teaching methods than the control group ($p=0.007$), and were also able to utilise better questioning skills than control group teachers ($p=0.003$).

Facilitator observations showed experimental group teachers generally asked mostly close-ended questions and one or two open-ended questions, while control group teachers only asked simple-recall questions or close-ended questions. Both experimental group teachers and control group teachers were, on average, able to give feedback to pupils about incorrect responses only, in a manner that encourages further effort.
Experimental group teachers used a significantly greater amount of learner grouping than control group teachers ($p=0.0003$). Generally, experimental group teachers used flexible groups without assigned roles, while control group teachers used the whole class only (no groups).

There was no gender discrimination between experimental and control group teachers as far as

---

[6] The "p" value is a measure of statistical significance. "p" values less than 0.05 are considered to be statistically significant at the 95 percent confidence level.

giving equal opportunities to boys and girls to participate in lessons (both groups scored a composite score of four in the classroom observation continua).

Table 8 shows that some differences were found between experimental group teachers in their teaching methodologies. EXP1 teachers used a significantly greater variety of teaching methods than EXP2 teachers (p=0.009) and EXP3 teachers (p=0.02) respectively. EXPI teachers were significantly better at questioning skills than EXP3 teachers (p=0.05), but there was no significant difference between EXPI and EXP2 teachers. This observation suggests that teachers having the most LP training (i.e. EXP1 has been in the project since the pilot) use a greater variety of teaching methods and better questioning skills. EXPI teachers were better than EXP3 teachers at providing feedback to students (p=0.009).

There was no significant difference between EXPI and EXP2 teachers or between EXPI and EXP3 teachers in their use of learner grouping; however, EXP2 teachers were better at grouping learners than EXP3 teachers (p=0.05).

Along each row, the underlined figure is significantly greater than the shaded figure(s) (p<0.05).

The majority of experimental group teachers interviewed reported that the ITEC training helped them improve in group teaching and pupil-centred learning.

3.3.2.1.3. Utilisation of Materials and Resources
Table 9 shows that ITEC-trained teachers utilised resources in the classrooms far more than the control group teachers.

Table 7: Summary of Composite Scores of Teaching Methodologies Between Experimental and Control Groups

	Experimental	Control
Use of Teaching Methods	3.42**	2.89
Questioning Skills	2.84**	2.00
Teacher Feedback to Pupils	2.89	3.00
Learner Grouping	2.79**	1.22

** Significantly higher mean score (p<0.05) between experimental and control group (based on one-tailed t-test at 95 percent confidence level).

Table 8: Summary of Composite Scores of Teaching Methodologies Between Experimental Groups

	EXP1	EXP2	EXP3
Use of Teaching Methods	3.75	3.28	3.25
Questioning Skills	3.17	2.77	2.62
Teacher Feedback to Pupils	3.50	2.86	2.46
Learner Grouping	2.75	3.21	2.38

Experimental Groups and Control Group Classroom Observation Continua
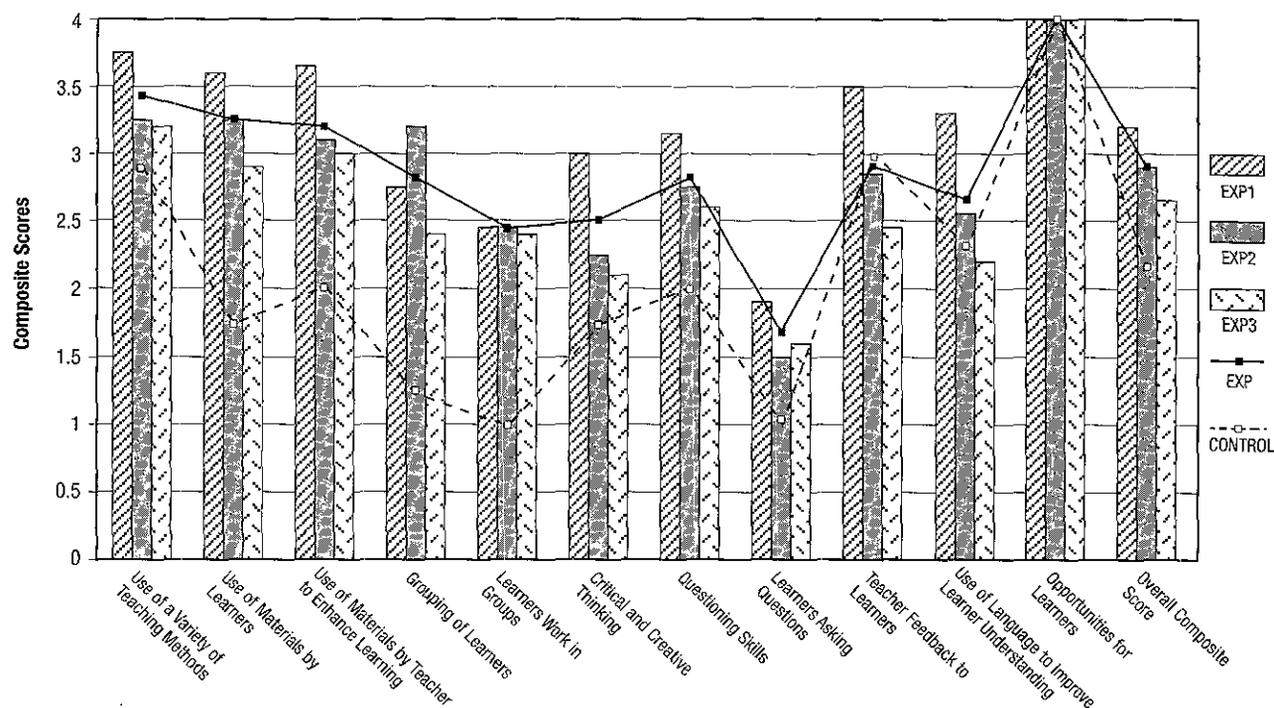
The classroom observation measured the extent to which teachers utilised materials to enhance learning (see Table 10). The use of materials was measured on a scale ranging from one (no materials) to four (more than two kinds of materials). Experimental group teachers utilised a significantly greater number of materials to enhance learning than control group teachers (p=0.0003). On average, experimental group teachers used at least two kinds of materials in a lesson to enhance learning, while control group teachers used only one kind of material. In the teacher questionnaires, experimental group teachers generally agreed that there were sufficient materials available, while control group teachers mentioned that such materials were generally insufficient in their classrooms.

Once again, length of time in the LP programme is associated with better observation scores (see Table 11). EXPI teachers used a significantly greater number of materials to enhance learning than EXP2 teachers (p=0.05). On average, EXP1 teachers used more materials than EXP3 teachers, but the difference was not statistically significant (p=0.052). There was no significant difference between the number of materials used by EXP2 and EXP3 teachers.

ITEC-trained teachers reported that ITEC materials were more relevant to their teaching than the previous materials they used in the classroom. They also reported that ITEC materials were suitable to both urban and rural pupils as well as for multiple subjects. One teacher mentioned that the ITEC materials assisted her pupils in self-discovery. Teachers said that ITEC had taught them how to improve their use of teaching aids. In particular, they indicated that they were able to make teaching aids:
• More relevant to the subject being taught;
• More accessible to pupils (e.g. placing them at pupil eye-level and allowing pupils to touch the teaching aids); and
• More attractive.

This observation indicates that ITEC has succeeded in its goal of producing materials which support and complement the syllabus.

3.3.2.2. Pupils
3.3.2.2.1. Utilisation of Materials and Resources The classroom observation measured the extent to which pupils utilised materials. The use of materials by pupils was measured on a scale ranging from one (none of the pupils manipulate materials) to four (learners share and all manipulate materials in groups or pairs).
Pupils in the experimental group made significantly greater use of materials than control group pupils (p<0.0001). On average, most experimental group pupils either shared or manipulated all materials, while control group pupils either did not manipulate materials at all, or some manipulated materials while others watched (see Table 12).

[7] Eighty percent of the experimental classrooms and 100 percent of the control classrooms did not have electricity.

Table 9: Resources Used in Observed Lessons

	Experimental (% of classrooms)	Control (% of classrooms)
Prescribed Textbooks	24.24	10.00
Exercise Books	68.57	40.00
Wall Charts	56.76	10.00
Chalkboard, Duster & Chalk	94.74	75.00
Power Points/ Electricity[7]	2.86	0.00
Visual Teaching Aids	61.11	22.22
Other Reading Materials	35.14	0.00
Self-Made Posters or Materials	65.79	0.00
Other Materials	48.15	14.29

Table 11: Composite Score of Utilisation of Materials to Enhance Learning Between Experimental Groups

	EXP1	EXP2	EXP3
Use of Materials to Enhance Learning	3.64	3.08	3.00

Along each row, the underlined figure is significantly greater than the shaded figure(s) (p<0.05).

Table 10: Composite Score of Utilisation of Materials to Enhance Learning Between Experimental and Control Groups

	Experimental	Control
Use of Materials to Enhance Learning	3.22**	2.00

** Significantly higher mean score (p<0.05) between experimental and control group (based on one-tailed t-test at 95 percent confidence level).

Table 12: Composite Score of Pupils Utilisation of Materials Between Experimental and Control Groups

	Experimental	Control
Pupils Utilisation of Materials	3.27**	1.75

** Significantly higher mean score (p<0.05) between experimental and control group (based on one-tailed t-test at 95 percent confidence level).

Pupils in EXPI made significantly greater use of materials than pupils in EXP3 (p=0.05), but there were no significant differences between other pairings of the three experimental groups (see Table 13).

Some of the experimental group teachers interviewed mentioned that they had given pupils the freedom for "hands-on" working with the materials whereas before they were only on display and out of reach of the pupils. It was also mentioned in the previous section that teachers were able to make support materials more accessible to pupils.

3.3.2.2.2 Pupil to Pupil Interaction
The classroom observation instrument measured the following aspects of pupil to pupil interaction:
• The extent of learner grouping in the classroom on a scale ranging from one (no grouping) to four (flexible groups and assigned roles);
• How pupils worked in groups on a scale ranging from one (learners sit in groups but work as individuals) to four (groups of learners discuss problems, questions and activities); and
• Critical and creative thinking activities on a scale ranging from one (teacher lectures, learners listen to teacher) to four (learners involved in discussions and problem solving and/ or creative activities).

Composite scores for each aspect were calculated by averaging the scores within the experimental and control groups (see Table 14).

Only two control group classrooms used learner grouping, so no statistical analysis was performed to compare the difference in composite score for this item.

Thirty-two of the 39 experimental group pupils (82 percent) and only two of the 10 control classrooms (20 percent) worked in groups. Experimental group pupils generally worked in groups where limited interaction occurred (not all the members in the group interacted). Experimental group pupils scored a higher composite value on the critical and creative thinking activities continuum than those in the control group (p=0.01). Generally, experimental group classes were equally divided between being involved in teacher-directed activities and in sharing of ideas amongst themselves. Control group pupils were generally involved in teacher-directed activities.

These results indicate that Module 1 on Learner Centred Education is at least partially successful. Interviews with key government informants showed that learner-centred education is a goal of the provincial department of education. Thus, the ITEC programme appears to be directly supporting this goal.
ITEC teachers' classes showed significantly greater abilities to solve problems and draw conclusions, whereas control pupils were utilising more rote-memorisation methods.

Recommendation: ITEC is moving towards its goal of learner-centred education, but still needs improvement in this area.

Table 15 shows that there were no significant differences between the three experimental groups on how pupils worked in groups. However, since EXP1 pupils scored a significantly higher composite value on the critical and creative thinking continuum than those in the other two experimental groups, teachers' length of time in LP has a positive effect on pupils' critical and creative thinking abilities. Generally, EXP1 pupils were involved in the sharing of ideas, while pupils in EXP2 and EXP3 were involved in more teacher-directed activities.

Table 13: Composite Score of Pupils Utilisation of Materials Between Experimental Groups

	EXP1	EXP2	EXP3
Use of Materials to Enhance Learning	3.58	3.29	2.91

Along each row, the underlined figure is significantly greater than the shaded figure(s) (p<0.05).

Table 14: Summary of Composite Scores of Aspects of Pupil to Pupil Interaction Between

	Experimental	Control
Extent of Learner Grouping	2.79**	1.22
How Learners Work in Groups[8]	2.45	1.00
Critical and Creative Thinking Activities	2.50**	1.78

Experimental and Control Groups

** Significantly higher mean score (p<0.05) between experimental and control group (based on one-tailed t-test at 95 percent confidence level).

Table 15: Summary of Composite Scores of Aspects of Pupil to Pupil Interaction Between Experimental Groups

	EXP1	EXP2	EXP3
Extent of Learner Grouping	2.75	3.21	2.38
How Learners Work in Groups	2.70	2.45	2.67
Critical and Creative Thinking Activities	3.00	2.31	2.23

Along each row, the underlined figure is significantly greater than the shaded figure(s) (p<0.05).

Recommendation: ITEC's estimation that teachers should have three years of INSET seems to be justified. Although it is evident that even one year of LP participation has created change, EXP2 teachers should be encouraged to stay in the programme.

3.3.2.2.3. Pupil to Teacher Interaction
The classroom observation measured two aspects of pupil to teacher interaction:
- How learners asked questions on a scale ranging from one (learners ask no questions) to four (learners ask questions which show creative thinking even without the teacher's encouragement); and
- Teacher feedback to learners on a scale ranging from one (gives no feedback/gives feedback in a manner that discourages further effort) to four (gives feedback about correct and incorrect responses in a manner that encourages further effort).

Table 16 shows that experimental group pupils were significantly better at asking teachers questions than control group pupils (p=0.03). However, on average, pupils in the experiment group asked simple questions only, while control group pupils asked no questions at all. There was no significant difference between the experimental and control groups on how teachers provided feedback to learners. On average, teachers in both groups provided feedback about incorrect responses only, but in a manner that encourages further effort.

Recommendation: ITEC workshops should place more emphasis on encouraging pupils to ask questions, and also endeavour to strengthen questioning skills amongst pupils.

Table 17 shows that there were no significant differences between the three experimental groups on how pupils asked teachers questions (on average pupils in each experimental group asked teachers simple questions). However, length of time in the LP programme is associated with differences in the teacher's ability to provide feedback to learners. Teachers in EXP1 were better at providing feedback to learners than those in EXP2 (p=0.05) and EXP3 (p=0.009). On average, EXPI teachers provided pupils with feedback about correct and incorrect responses, while EXP2 and EXP3 teachers provided feedback only about incorrect responses.

One experimental group teacher interviewed by the evaluation team mentioned that pupils at her farm school have gained confidence in their communication skills to the extent that some of the pupils were confident enough to talk to the farm owner's wife!

3.3.2.3. Classroom
Overall, the majority of experimental group and control group classrooms had adequate seating and writing surfaces for all pupils, a chair and table for the teacher, adequate lighting, adequate space for movement between desks, comfortable ventilation and temperature, a cheerful environment, cemented or tiled floors, and no disruptive outside noise. Recommendation: Given the relatively good state of classrooms in the experimental and control groups, ITEC should consider implementing the LP in schools with poorer infrastructures to see if LP will work in such schools.

3.3.2.3.1. Display of Materials
As mentioned earlier in this section, experimental group teachers used a significantly greater number of teaching aids and materials than control group teachers. Experimental group teachers noted in teacher interviews that the ITEC training had assisted them to make more relevant materials and to make those materials more accessible to pupils.

Table 16: Summary of Composite Scores of Aspects of Pupil to Teacher Interaction Between Experimental and Control Groups

	Experimental	Control
Learners Asking Questions	1.64**	1.00
Teacher Feedback to Learners	2.92	3.00

** Significantly higher mean score (p<0.05) between experimental and control group (based on one-tailed t-test at 95 percent confidence level).

Table 17: Summary of Composite Scores of Aspects of Pupil to Teacher Interaction Between Experimental Groups

	EXP1	EXP2	EXP3
Learners Asking Questions	1.91	1.50	1.54
Teacher Feedback to Learners	3.50	2.86	2.46

Along each row, the underlined figure is significantly greater than the shaded figure(s) (p<0.05).

3.3.2.3.2. Classroom Arrangement

Both experimental and control group teachers generally agreed that their class sizes were not too large (see Table 18). Although the teacher questionnaires on average indicated that both experimental and control group teachers utilised group work activities, the classroom observation continua clearly indicated that experimental group pupils worked in groups while control group pupils did not. There were no clear cut indicators from the data collected which compared whether actual classroom arrangement was in rows or clusters. Classroom sizes, on average, are less than the 40 students per classroom proposed by the Education White Paper.

3.3.3. Pupil Performance

Table 19 presents a comparative summary of mean mathematics and language assessment scores between the experimental groups and the control group

As shown in Table 19, experimental students outperformed control students on maths tests, except for Sub-B. This demonstrates that participation in the ITEC programme is significantly and positively associated with better maths performance. This pattern is consistent at the individual grade levels, except for Sub-B. The lack of a significant difference between experimental and control scores in Sub-B is explained by the absence of the EXP1 group in this class which may have "pulled" up the average for experimental maths scores in this class.

Language scores show a different pattern, however. There is little difference between the experimental and control groups, and where a difference occurs (as in Sub-B and Std 1) there is no consistent pattern. This can be partially explained by the fact that a revised ITEC language module was only introduced to participants in the last eight months, with training completed only in August 1995, a few days prior to this evaluation. This suggests that there has not been enough time to implement the revised language module. Recommendation: ITEC has been providing materials for language training since its inception, and the analysis showed little difference between experimental and control groups; therefore, this is an area which ITEC must improve and monitor closely.

The pupil assessment activity results suggest that any changes in student performance as a result of involvement with ITEC is highly dependent on length of time the programme has been implemented in the classroom. This finding is further confirmed by looking more closely at the differences between the three experimental groups.

Table 20 presents an analysis of the differences between the individual experimental groups for maths and language scores. There are a few discerning trends. First of all, for all math:; results, there is a significant improvement in mean scores with longer involvement in the ITEC programme (p<.001). Specifically, EXP1 students outperformed all other groups (including control) while EXP3 students show very little difference in their scores from controls or EXP2.

[9] This farm school classroom had a total of four pupils of whom two were Std 2 pupils whose data was not analysed in this evaluation.

Table 18: Experimental and Control Group Classroom Sizes

	Minimum	Maximum	Average
Experimental	2[9]	51	30.3
EXP1	2	51	31.2
EXP2	15	49	31.2
EXP3	13	50	33.5
Control	22	42	30.6

Table 19: Mean Assessment Scores

	MEAN SCORE (% correct answers)				
	EXP1 N=343	EXP2 N=437	EXP3 N=402	TOTAL EXPERIMENTAL N=1182	CONTROL N=306
All Students					
Maths	75.6	67.8	65.9	69.4**	64.5
Language	52.0	50.0	56.0	52.6	51.5
Sub-A					
Maths	78.8	75.7	60.0	78.3**	71.3
Language	54.9	56.4	56.7	55.4	56.0
Sub-B					
Maths	-	73.7	69.3	71.3	69.8
Language	-	63.7	68.7	66.3**	46.6
Std 1					
Maths	61.0	63.1	64.7	63.4**	50.0
Language	38.1	41.7	44.6	42.3	48.7**

** Significantly higher mean score (p<.001) between total experimental and control group (based on one-tailed t-test at 95 percent confidence level).

Secondly, between experimental groups, the language scores show little difference in student performance despite length of involvement in the programme. As explained above, this may be due to the fact that a revised ITEC language module was only recently introduced to the programme and has not been implemented for sufficient time to produce improvements in performance. If this is the case, ITEC's success may therefore be highly dependent on the time teachers have been enrolled in the programme, with longer classroom and teacher participation associated with better student performance.

3.4. INTERVIEWS AND DISCUSSION
3.4.1. Accreditation

During discussions with key informants, ITEC staff and teachers, the issue of accreditation consistently was raised, with many believing that the Education Department recognised accreditation as a critical incentive for teachers. Moreover, teachers' responses indicated a desire for accreditation. Department officials felt that accreditation would strengthen ITEC's programme. Rubasana Teacher Training College indicated that it was interested in working cooperatively with ITEC in training junior primary PRESET students. Likewise, a key informant involved in the Fort Hare/ South Australia University programme indicated that that institution is developing a distance education PRESET/ INSET programme to upgrade teachers' qualifications in the Eastern Cape. She stated that NGOs such as ITEC will be asked to submit their programmes for accreditation and distance delivery in 1996, for programmes to begin in 1997. While several respondents did not respond to questions about highest academic and professional qualification, Table 21 and Table 22 illustrate the responses of teachers who provided information relating to the academic and professional qualifications of control and ITEC teachers.

These tables indicate only 73 percent of ITEC teachers have successfully completed secondary school, and 68 percent of their highest training is a Primary Teaching Certificate. Given the disparity of academic and professional qualifications, ITEC will need to carefully analyse the level of accreditation required by teachers at the diploma level. Recognition of prior learning as encapsulated in the National Qualifications Framework should also be taken into consideration when identifying the level of accreditation.

3.4.2. Partnerships
Nearly all government personnel interviewed for this evaluation indicated that the government is interested in forging partnerships with NGOs such as ITEC. They felt that NGOs and government

Table 20: Maths and Language Assessment Activity
Results Comparison Between
Experimental Groups

Maths Assessment Activity

	EXP2 (67.83%)[1]	EXP3 (65.97%)	CONTROL (64.54%)
EXP1 (75.62%)	YES[2]	YES	YES
EXP2 (67.83%)	-	NO	YES
EXP3 (65.97%)	-	-	NO
EXP (69.46%)	-	-	YES

Language Assessment Activity

	EXP2 (50.02%)	EXP3 (56.07%)	CONTROL (51.54%)
EXP1 (52.07%)	NO	YES	NO
EXP2 (50.02%)	-	YES	NO
EXP3 (56.07%)	-	-	YES
EXP (52.67%)	-	-	NO

Table 21: Academic Qualifications of Teachers

	Experimental Group 1 N = 10	Experimental Group 2 N = 12	Experimental Group 3 N = 8	Total Experimental Group	Control Group N = 6
Not matriculated	40%	17%	25%	27%	0%
Matriculated	60%	66%	75%	66%	100%
Bachelor's Degree	0%	17%	0%	7%	0%
Total	100%	100%	100%	100%	100%

Table 22: Professional Qualifications of Teachers

	Experimental Group 1 N = 12	Experimental Group 2 N = 13	Experimental Group 3 N = 9	Total Experimental Group	Control Group N = 9
PTC	67%	54%	89%	68%	44%
PTD	33%	31%	11%	26%	44%
JPTD	0%	15%	0%	6%	12%
Total	100%	100%	100%	100%	100%

[1] Mean assessment activity results are in parentheses.
[2] "YES" indicates a (one-tailed) significant difference between mean scores; "NO" indicates no significant difference.

could work together to successfully fund-raise for programmes which would build government capacity to deliver quality education. The director of Early Childhood Development which is responsible for the educational requirements of children from birth to nine years old, stated that she had already put into place a partnership with ITEC. ITEC strategic planning documents indicate that ITEC would like to assist the government through training trainers, and by addressing the needs of teachers through research and development.

3.4.3. Management Information Systems

The evaluation team leader met with the ITEC information technology consultant retained to develop a management information system (MIS) for the LP programme. It is clear that ITEC has gathered fairly complete information on participants and some information on the school context. At the time of this evaluation, the consultant was beginning to develop programmes in response to needs of ITEC to monitor fee payment, completion of modules, and other information. During the discussion involving the consultant, ITEC LP personnel and the evaluation team leader, it was recommended that ITEC utilise and adapt the observation continua, pupil assessment exercises, and other instruments during field visits to monitor increasing teacher competencies. A concern was raised by ITEC about the added expense of the system. However, as effective impact monitoring is a responsibility of all NGOS, Khulisa recommends that it be considered art overhead expense. Other indicators of success should also be monitored, with a particular focus on systemic issues such as:

- Reducing the level of repetition and drop-out in the first few years of schooling which, on a national level, is estimated at 30 percent;
- Increasing teacher-pupil contact time which, due to teacher absenteeism and releasing children after examinations (rather than the continuation of teaching), was estimated by one informant at 60 percent of the required time; and
- Monitoring the promotion statistics of ITEC participants to determine the promotion rate of ITEC teachers.

3.5 RECOMMENDATIONS

The evaluation team has the following recommendations for ITEC's Lower Primary programme.

- While ITEC teachers showed that they were better at learner-centred teaching than control teachers, problems with learners in groups and learners asking questions indicate that ITEC should work with teachers to improve the performance of learner-centred teaching.
- Monitoring should focus on competencies derived from the eight modules so that the language problem is not repeated in the assessment results. There may be a need to redesign the language section and work more with the teachers. It is important that ITEC share the test data with the teachers.
- ITEC's estimation that teachers should have three years of INSET seems to be justified. Although it is evident that even one year of LP attendance has created change, EXP2 teachers should be encouraged to stay in the programme.
- ITEC workshops should emphasise allowing pupils to ask questions and strengthening these skills amongst pupils.
- ITEC should implement the LP in schools with poorer infrastructures to test if the programme is successful in such schools.
- ITEC has been providing materials for language training since its inception, and the analysis showed little difference between experimental and control groups; therefore, this is an area which ITEC must improve and monitor closely. • Accreditation is critical from both government and teacher perspectives. ITEC should pursue integration in PRESET with Rubasana, and with INSET through the Fort Hare / South Australian Distance Education Initiative, Rhodes, or an outside agency. ITEC could provide the practical competencies while the tertiary institution would be responsible for the theoretical competencies. ITEC should carefully consider the level of accreditation based on additional research and the academic/ professional qualifications of its participants.
- ITEC should continue to strengthen partnerships, especially in the language modules, to convince government to implement the ITEC approach. ITEC should follow its strategic plan to shift its focus to research and development and training of trainers, rather than implementing the LP with ever greater numbers in the whole province.
- ITEC is beginning to collect data for a management information system. ITEC facilitators should use the classroom observation instruments and pupil assessment exercises to continuously monitor the progress of teachers and pupils. ITEC should also monitor the effect of its programmes on pupil drop-out rates and repetition rates and other similar pervasive problems in the province.

SUMMARY AND CONCLUSION: "PUTTING 'THE STUFF IN PLACE"

Professor Tony Morphet, University of Cape Town

My task as I understand it, is to a put together an account of today's conference proceedings and for my title I have borrowed a phrase from Noleen's commentary on the ITEC case study. She spoke of the process of constructing the evaluation of ITEC as 'putting the stuff in place', and broadly speaking, I think that is what we have been working at all day. The 'stuff' is the activity of evaluation.

It struck me while she and Jennifer were talking, and as I was preparing this, that we have been on a trip. It's been a journey which has been quite difficult for most of us. The general direction has not been a problem since we got the compass bearing on evaluation from Nick Taylor at the start - but it hasn't been that easy to follow the detailed twists and turns in the track. To be candid it quite often seemed as if the roads were flooded, the signposts down and the bridges under water - the effect, I guess, of the high-level academic abstractions washing over our old familiar ground. If I can borrow another lovely phrase from Noleen - the 'High George' has given us a bit of difficulty in finding our way. However, it has been an important and absolutely necessary trip. We have had to leave behind a home base on which we used to feel comfortable and to set out to try to find some new, firmer, ground. The reason is simple. The trust, confidence and legitimacy that we have based our work on for the last decade at least, has collapsed. Nick Taylor gave us the terms of the collapse right at the start - evaluations of NGOs' work have not been able to measure the effects of INSET in such a way that pupils can be shown to have benefited. The JET review of the INSET evaluations has shown that the evaluation; are not able to provide evidence that money being spent produces a discernible effect on the pupils. It is no longer possible to give an adequate public account of either the money being spent or the work being done. Jonathan Jansen, speaking from his own extensive experience as an evaluator, reaffirmed again and again this point made about the methodologies and findings of evaluations.

The question which the case study took up was whether evaluation practice might be able to help in establishing some new grounds for confidence - and if so what would it have to do? The three speakers were agreed on one thing - properly conducted evaluations would certainly be a means of working towards a new level of trust and confidence, but they weren't in agreement about what a proper evaluation actually is. Johann Louw raised the banner of science. He split programme research from programme evaluation in order to focus as clearly as possible on the benefits of methodologically rigorous scientific studies of programme interventions. He suggested quasi-experimental research designs to provide not only comparability across programmes in terms of objective results, but also to make progress towards the definition of causal relationships between inputs and impacts. The model was as pure as possible and undoubtedly powerful but it seemed a little more difficult to get it to work in the messiness of NGO field activity. Johann even had a problem fitting the description into his time slot. Sometimes it wasn't clear whether he was trying to bring INSET work into the laboratory or to take the lab' into the field. Science has enormous appeal but it isn't always easy to take it out on the dirt roads. Joe Muller acknowledged the value and capacity of science but he had another

evaluation model as well - the expert judges panel. He split the truth of the scholar /scientist from the more worldly truths of fairness and justice, arguing that both were important but because they came from people with different kinds of training it would be really hard to find someone who could do both well. The pure and precise world of the truth-seeking scholar would always be fairly remote from the public world of power which the adjudicator normally inhabits.

Jonathan Jansen, hiding his scholarly credentials just a little, spoke as a professional evaluator - a person whose first responsibility was to make judgements that would be sound enough to assist both programme people and their donors to make real-time judgements about programme action and its costs. He was critical enough of current check point evaluations ('drop-in visits which took the programme statements at face value) but he wasn't in favour of full-on quasi-experimental causal explanations either. He put the case for longer, deeper, 'clinical-type' studies. He seemed to be connecting himself to both the forms of truth which Joe had described.

At the end of the first session, I thought the case for evaluation was still very much open. There had been three different kinds of bid for a procedure of sound judgement (sound enough, that is, to re-establish the necessary trust, confidence and legitimacy in the business of NGO INSET.) Neither science (with a capital S) nor the logics of power (with or without the capital P) nor yet Jansen's even-handed 'clinical' pragmatism had proved persuasive enough to settle the specific procedures to take us to the firmer ground. In fact during the discussions I thought there were several places where we only narrowly avoided serious skids back into the old deep ruts in which NGOs and their donors have become accustomed to slinging mud at each other. In other words the levels of audience trust and confidence were sometimes low indeed, but they weren't broken and there was (at the tea break) still enough firm road to move forward.

Peter Weingart gave us, after tea, an experience which, to us in the educational community, is altogether too rare. His account of the falling levels of confidence and trust in the great Anglo-US scientific community - of the legitimacy crisis within science itself - gave us a comparative example which refracted our own local experience with a diamond-like clarity and brilliance. Within minutes of his opening, the parallel forms (if not the dimensions) of the issues of legitimacy became clear. Science itself, like NGO INSET work, had lost the established basis of its own public justification and had been obliged to set out to establish new valid forms in which to render its account of its practices to the public. His paper provided an even greater critical thrust through his description of just what measures the scientific community had taken to re-establish public trust in the scientific enterprise. At the heart of these was effective evaluation.

Weingart grasped the sharpest nettle of all when he showed how it was impossible to evaluate science according to any criteria which were not from science itself - but how equally impossible it was to allow scientists simply to evaluate themselves according to their own criteria. The absolute requirement for any evaluation which would be able to sustain both public trust and the confidence of the scientific community, was that the judgements had to be made in terms of measurable (and quantifiable) indicators which would be acceptable both to the scientists on the inside, and to the public on the outside, of the scientific community.

Thus the heart of the matter, we came to see, was the construction of valid indicators of performance.

Weingart described how the construction of indicators depends on plausible theories of representation. Any theory of representation, in his view, generates hypotheses about the relationships between complex processes and outcomes on the one hand, and visible, measurable, events on the other. The indicator is the place where the complex, often invisible, process (say of learning/ teaching) re-presents itself for observation. Weingart's scientific example gave him the means of showing just how these central issues of evaluation had been worked through in the scientific community.

It was this account that turned the light back on to the position that the NGO community is presently in. - and illuminated the track ahead. The legitimacy road we are travelling is from one theory of representation to another. NGOs used to be legitimate because they had a plausible theory o£ representation - and it provided an agreed set of indicators on which judgements could be made. This theory of representation, never very explicit but firmly in place throughout the 80s, was a politically generated theory which was acceptable to foreign governments, local donors, NGOs and clients alike. What made it plausible were shared interpretations of the political conditions of the struggle in education and elsewhere. Its terms were that good educational intervention work would be re-presented in the forms of pedagogical relationships. Black teachers were the target community and the ideal relationship form was group-based, participatory, supportive, critically reflective collaboration - in a word. 'democratic'. The goal was the building of a new educational and political culture through changing the self-awareness of teachers.

The indicators for an NGO's performance that this theory generated are familiar to all of us. They were the visible signs that the NGO would be able to contribute to the 'new educational culture' - that it was working in the right place, with the right kinds of teachers; that it was under the direction (or more usually co-ordination) of a black person, that its own practices were adequately non-._ hierarchical, non-racist, non-sexist, democratic and participatory; that it was accountable to 'the community'; and that it had an adequate representation of credible Black leadership on its Trustee Board - and so on.

This is what has collapsed - and for several reasons. Among the more important are that NGOs no longer have a publicly justifiable role in building a new educational/political community, but perhaps the most significant and interesting reason is in the fact that the 'progressive' teacher culture is now the norm - the battle has been won. No!., it seems, that it has helped the pupils in the ways we thought it would.

So, with our traditional legitimacy gone, and our theory of representation unable to give us valid indicators, the question arises 'How does one construct a new and better theory?'. Peter was asked this question and in answer he gave two somewhat different responses. The first was that it is built through the research process - the reference was to Johann's account of the continuous feedback loop running between measured programme performance and improved programme design. The second, rather different answer was through negotiation between the stakeholders - this I thought referred back to Joe's account of the way things are done in the public world of power.

So at the end of Peter's marvellous contribution we returned, vastly more enlightened and focused, to a position quite close to the one we had been in at the end of the panel discussion. One means of building a theory of representation is public and political - the other is 'private' and academic. The question was still open - how do the different pieces get put together. Who does what on the ground? How do we 'get the stuff in place?' The decision to present the ITEC case after lunch was inspired - principally because it was able to take us to the field of practice and show what it means 'to put the stuff in place'. The presentation achieved many things. It gave the evidence to substantiate a claim that NGOs can in fact have an impact on pupils, it showed how an effective, credible evaluation goes about its work; it demonstrated the implementation of the kind of quasi-experimental design that Johann had described - and at the same time it showed the limits that actual field conditions impose on the search for the objective truths of the scientist - and what such work costs! But the important thing which it showed was the detailed process of negotiation between the stakeholders which produced the indicators on which the quantitative measures could be made. How broad that process was, was itself instructive - initiated by Penny, Jennifer and Noleen, it was extended to include the evaluation field worker trained by Jennifer, to the teachers themselves and eventually through Noleen's planned 'big indaba' to the teachers' union and the community at large.

What was going on in the negotiation was the construction of a plausible theory of representation, and, just as in Peter's scientific case, the indicators had to be acceptable to Noleen and her colleagues working in the NGO in the field, and to those in JET who were accountable for the use of public money. Moreover the theory had to be able to represent the complex sets of relationships between the pupils in the classrooms, their teachers and the INSET team in ways which would make the transmission of capacity visible and measurable. So they finally got to the tests and scores and specified behaviours which Jennifer set out to measure under controlled conditions.

The limits of what she was able to do became clear at several points in her presentation and she summed these problems up in repeating the anguished phrase 'no baseline data'. What she was saying in effect was "let's agree, at the start of a project, on what will count as valid indicators, then we will be able to take one measure at the beginning and another at the end and we will be able to demonstrate the effects of the intervention with confidence". No-one asked Jennifer, or anyone else for that matter, why there was 'no baseline data' but it would have been an important question to ponder because it would have led us back towards Peter's other description of the ways in which indicators are constructed - namely in the research process - and that in turn would have taken us back to very beginning to Johann s science. Negotiation can take us only so far in putting the stuff in place - what we need in addition is a body of tested public knowledge about programme design that will make reliable indicators available at the beginning of an intervention - and that knowledge can only come from the kind of research work which Johann described so carefully.

Thus what the case study was able to show the rest of us was what the new theory of representation could look like - and what it might mean for NGO activity. It also showed us how to begin the work.

The capacity to do this work is already present. What is required is an extension throughout the NGO field of the form of negotiation which the ITEC evaluation displayed and that needs to be backed by consistent academic research into programme design. Donors, NGOs, academics, teachers and field staff each need to contribute to the construction of an agreed theory of what will be used to represent success. It is only on such a secure and tested base that NGOs will be able to settle once again on stable and legitimate ground.

JET has an obvious and important role to play in the work, but not to be overlooked is the crucial work that it has already done. The planners are to be congratulated on the superb design of the conference which in a single day took evaluation work and NGO practice along the difficult path into a whole new territory.

APPENDIX 1

INSET, NGOs AND EVALUATION: A REVIEW

Presented by Nick Taylor
to the Kenton-at-Settlers Conference, 27-30 October 1995

1. INTRODUCTION

This review is part of the NGO component of the Teacher Education Audit commissioned by the national Department of Education in April and funded by Danida. While the bulk of the NGO report, researched and written by Penny Vinjevold, focuses on a quantitative delineation of NGOs involved in INSET, the present paper examines the quality of their work.

A total of 54 evaluations covering 33 INSET programmes were reviewed. These were gathered by requesting all NGOs who participated in the Audit to supply copies of any evaluations conducted in the last four years. In addition, Khulisa Management Services were commissioned to gather evaluation studies from funders as part of their financial survey of funders. The evaluations surveyed are referenced in the text in italics and listed in Appendix 1a. All other works referred to in the present chapter are listed in the Bibliography. The works reviewed here cover almost exactly one-third of all NGOs identified in the main report as undertaking INSET (99 in total), and just over a quarter of the 196 programmes offered by these providers.

Conventional wisdom has it that NGOs, being small and flexible, close to their client communities and free from the clumsy bureaucracy of government, are able to deliver services more efficiently and effectively than the state. Korten (1990) recognises three ideal phases in the development of an NGO:

♦ The organisation focuses on a specific local problem and develops an **innovative** solution

♦ After achieving success, the project gears up its activities and begins to implement the model on a larger **scale**

♦ A mature phase is reached when the NGO becomes involved in **policy advocacy**, thus making the lessons learned available to government and other actors

The present paper examines these claims with respect to INSET NGOs, through an analysis of the evaluations under review. There is at present in South Africa a public campaign drawing attention to the fact that foreign donors are redirecting resources away from NGOs, and calling for the allocation of RDP and other funds to these organisations. (See, for example, IDT, 1995). A major problem bedevilling this debate is that, by and large, NGOs have slipped into Korten's third phase - policy advocacy - without producing cogent evidence that their innovations are making a difference and that they can deliver to scale on a cost-effective basis.

Until they can demonstrate delivery there is no case, other than blind faith, for prioritising the spending of public and private resources on NGO activities such as in-service teacher training. Improving the quality of schooling certainly must rank as a priority if South Africa is to develop a skilled and educated population. However, there are far too many other needs, like water and housing, for any sector to have an automatic call on funds. For example, as part of his evaluation of the Palabora INSET programme, *Schollar* undertook a survey of priorities amongst the community served by the project. The respondents exhibited a high level of awareness and approval of the project (40%). It is significant, however, that INSET programmes for teachers (at 23,1%) rated only seventh on the list of development priorities, after employment (75,6%), literacy (72,6%), post-school bursaries (62,6%), services (water, electricity) (54,6%), pre-school provision (50,3%) and roads (23,6%).

In the past NGOs were funded because they occupied the anti-apartheid moral high ground. This ground is rapidly crumbling and their only hope of survival is to construct a well argued case about the quality of their service delivery. The present review is driven by this perspective.

2. ISSUES IN EVALUATION METHOD

2.1 The qualitative/quantitative standoff: finding a suitable metaphor

A discussion on the state of evaluation practice in South Africa, as revealed by the work under review, is necessary in order to sift the wheat of the findings of these studies from the chaff. And the first issue that requires a closer look is the paralysing standoff between 'qualitative' and 'quantitative' approaches to evaluation. Two examples illustrate the respective strengths and shortcomings of the two paradigms.

Schollar's evaluation of the Palabora project is a classic case of a qualitative study, the heart of which consists of a set of sit-in visits to the classrooms of project teachers. *Schollar's* sure-footed classroom observations provide a wealth of insight and information about schools, teachers, pupils and child-centred teaching methods. However, he concedes that only a quasi-experimental approach is capable of establishing whether or not the project intervention has had an effect on the learning outcomes of the pupils taught by project teachers.

Dyrenfurth's rigorous quantitative study of Protec, on the other hand, allows very little room for doubt that the project pupils have learnt more about technology than their non-project counterparts. However, he is candid about the weakness of the classroom component of his study, and there is a yawning gap between the project managers and their curriculum, on the one hand, and the improved knowledge of the pupils, on the other. Were the learning gains due to good, teacher-proof materials? Or were the teachers central in facilitating the pupil learning gains? What methods did the teachers employ? How were they trained?

Whether they work in the qualitative or quantitative paradigm it is highly unusual for evaluators to acknowledge the value of the rival approach in the way that *Schollar* and *Dyrenfurth* have done. Proponents of the respective approaches normally regard each other with deep suspicion, if not open hostility. We would argue that the two examples quoted above illustrate that neither

approach is complete, and that a combination of methods is needed. This is not to deny the existence of tricky epistemological questions concerning the nature of the real, its relation to description, and the position of the scribe. It is rather to assert the need - within a sensitivity to these issues - to use all appropriate means to gather and analyse empirical data in order to establish whether or not a particular project works and, if so, to disentangle the various factors which contribute to making it effective.

Quantitative methods have been criticised on the grounds of treating human subjects like the inanimate objects of physics experiments, or like rats undergoing behavioural conditioning. Extreme forms of the quantitative paradigm certainly used to, and occasionally possibly still do, rest on such assumptions. But, by and large, today such criticisms merely provide straw windmills for indignant qualitativists to tilt at.

A more appropriate metaphor, according to the latter, is to think of evaluation as being like literary criticism, or judging a work of art. *Bateson (1994)* describes the approach as follows:

> "Rather than being based on 'objective' quantifiable indicators, the evidence of [the project's] success should be documented collections of carefully selected anecdotes and testimonials, both positive and negative, couched in the professional judgements of qualified evaluators"
>
> *(ibid, 1994)*

It is this particular form of the qualitative approach which is most commonly employed in the evaluations under review. The anecdotes appearing in these studies are, in the main, not used as evidence by any stretch of the term, but as illustrations to the assertions of the evaluator. In many cases the anecdotes are not even quoted but merely referred to.

The test case for this methodology arises when two evaluators reach different conclusions from the same situation. For example, in their evaluation of the TOPS Management Programme, *Gilmour and Soudien* concluded "from the testimony of facilitators

and participants and from the physical organisation of the schools visited" that "Most schools from which participants are drawn are now running far more effectively than before" *(13)*, and that these changes are due to the "considerable impact from the TOPS programme" *(22)*.

Yet, in a review of existing data on TOPS, which included the above study, the *IEQ* evaluation notes gaps with respect to observational data and a lack of adequate comparison which would allow attributing change to the TOPS programme. These observations pose a grave challenge to *Gilmour and Soudien's* conclusions. What does the reader make of this contradiction? Because *Gilmour and Soudien* do not build a systematic case there is no way of testing their evidence and hence of adjudicating between the two reports.

Our own starting point is that the hard science approach employed in extreme forms of quantitative evaluations, and the art criticism metaphor used, at least implicitly, by the majority of the studies under review are equally inappropriate. We argue that evaluating the worth of an educational intervention such as an INSET programme is akin to a legal hearing in which a case needs to be built using all available evidence, argument is led and substantiated by skilled advocates, and judgement delivered according to established rules of evidence, balances of probability and respect for the rights of the evaluatees.

During this process the evaluator is not an objective observer: she can neither stand outside the project without altering its course nor remain indifferent to the outcome of the judgement. And the judge - the reader of the evaluation report - is inevitably swayed by the eloquence of the various advocates, but in the end it is the strength of the case, argued from the evidence to hand, which tips the scales one way or the other. A critical mistake made by many of the evaluators under review is to confuse their recommendations with the final judgement. Consequently, they don't see fit to fully outline the arguments leading to their conclusions, or to adequately substantiate them by means of evidence. In short, the case is incompletely made and the

recommendations often represent no more than assertions which require corroboration before they can acquire the status of solid pointers for decision making. The remainder of the present section is concerned with a discussion of ways in which the studies under review gather, sift, weigh, test and present evidence in order to build their cases.

2.2 Description of the project and its work
Many of the studies do not contain a description of the overall dimensions of the project under assessment. In our view, such a description is essential in order to orient and locate the reader. This description should include:

The objectives and intended outcomes: what the project is attempting to do. It is impossible to attach a value to any initiative if its intentions are unknown. A vital step in understanding the intentions of an intervention is a clear statement of the problem it sets out to address.

It is fashionable for INSET programmes to distance themselves from 'deficit models'. These programmes motivate their intentions instead in terms of a 'developmental' approach. However there is no getting around the fact that programmes that provide support, development or training to teachers consider that some teachers and schools are more in need of development than others. What is this need, and how does it manifest itself? are questions which are prerequisite to identifying suitable measures of success.

If, for example, a project is directed towards management training for principals, then it would seem logical for the evaluation to ask about the conditions in the schools which motivate for such training. Is it because there is no timetable after three weeks of the start of the year, because teachers come and go as they please, or because the school day is subject to frequent unscheduled interruptions? And if these are problems, then surely a measure of success of the project would be to look at the extent to which it addresses these dimensions of school management.

Evaluators who adopt a 'qualitative' stance by and large avoid this kind of logic, arguing that an obsessive focus on outcomes obscures

much that is valuable. The deficiencies identified above in *Dyrenfurth's* evaluation of Protec could be taken to illustrate their argument, but surely the point is not that evaluations should throw out the outcomes baby because of potential distortions, but should rather add components which render the bathwater more transparent.

The activities undertaken in pursuit of the objectives: what the project does. For example, four-day workshops, one-day planning sessions, classroom visits. Many evaluators either omit this aspect altogether or allow some but usually not all of this information to seep out incidentally during the course of the report. It is bewildering for the reader to be plunged into analyses of isolated activities without having an overall picture.

The scope of these activities. For example: how many teachers from which schools attended workshops of each type and how frequently. This is the most frequently omitted element of evaluations. In the overwhelming majority of cases the information is not available from the organisations under review. *Khulisa* and *Schollar* had to piece together what they could, from available project documentation and the results are far from satisfactory.

2.3 Evaluation questions

Evaluations may be commissioned by a variety of agents: the project management, its parent organisation (eg. a university), the board of trustees, donors, to name but the most common. In addition, agents may have a variety of reasons for the evaluation: donors are likely to be interested in fiscal accountability, while trustees may be more interested in management issues. Weiss (1990) points out that this complexity can result in the evaluation not being focused enough to satisfy any of the parties. She recommends that the questions driving the study should be formulated in consultation with major stakeholders and clearly articulated.

In Leat's (1990) view, stakeholders are often not sufficiently knowledgeable to ask appropriate questions or to assess the adequacy of the answers. This is especially true in what she refers to as programme accountability: assessing the quality and

relevance of the services offered by the project. This is the most difficult and hence the most underevaluated aspect of NGO work. Because of this problem, there is often a gap between the issues actually addressed by an evaluation and the importance attached to its findings: an assessment which reports favourably on management practices may be taken to imply that the services delivered by the evaluatee are of a satisfactory quality.

These observations are in accord with the findings of our survey. The majority of the evaluations under review do not deal anywhere near adequately with programme effects: the impact of the project on the quality of teaching and learning. And yet, more often than not, whether implicitly or explicitly, the findings are taken as a stamp of assurance of programme quality.

2.4 Gathering information

Four categories of data were used in the studies under discussion.

Self report data. This is far and away the most common type of information used in 'qualitative' studies. It is gathered by means of questionnaires and interviews. Questionnaires and structured interviews are amenable to a collation and quantification of answers, and hence deriving a weighted picture of stakeholder views. Unstructured situations, on the other hand, give more scope for views not considered by the project or evaluator and may be very useful in questioning fundamental assumptions.

Self report data is particularly useful in ascertaining the disposition of participants towards the programme. This information is important, but *Schollar's* study indicates that feeling good about a project, even being convinced that it is making a qualitative difference, is not sufficient evidence of such change.

In the assessment of outcomes, self report data must be subject to rigorous testing, not so much because of possible bad faith on the part of respondents, although this must remain a possibility, given the fact that evaluations often influence life and death decisions around the allocation of resources. Rather,

however, self report data must be tested for the influence of excessive good faith: programme staff are idealistic and dedicated; participants have worked hard and are excited about the possibilities of the new methods solving what seem to be insurmountable problems; everyone wants the programme to succeed. Under such circumstances it is easy to confuse best intentions, and changes in the outer forms of teaching and learning, with qualitative improvements.

Project documents. These are important, together with interviews with project staff, for understanding the goals, activities, intended outcomes and scope of the programme.

Classroom observations. Classroom observations provide an enormously rich source of data about general conditions in schools, qualitative assessment of teaching methods, the quality of learning, the use of equipment and materials, and on discrepancies between the forms of teaching and learning behaviours and their outcomes.

Expert judgement is indispensable in the assessment of these aspects of programme quality. But, as discussed in 2.1 above, the case needs to be systematically built before judgement is delivered. Obtaining commensurability of classroom assessments across different situations and observers, and making the criteria for such assessments transparent is a notoriously slippery

problem. Schedules for guiding lesson observations were used in a number of the evaluations reviewed, as a way of addressing this issue. *Khulisa*, for example, devised a schedule for quantifying and comparing certain classroom behaviours, such as the level and quality of pupil-pupil and pupil-teacher interactions.

Instruments for assessing cognitive skills. The most reliable instruments for this purpose are those which have been validated and used by other evaluators. This is the route taken by *Dyrenfurth*. However, when a project challenges the conventional curriculum (a situation which applies to most programmes covered by this review), by definition, no validated instruments exist. *Khulisa* obviated this problem by constructing a test in consultation with the evaluatees, while *Bateson (1995)* adapted an existing validated instrument with the help of the programme implementers.

Khulisa added a further component: individual pupil interviews in which the pupils were asked to solve given problems and to talk about their strategies. This provided a very fruitful source of insight.

2.5 Assessing effects
Which effects; what measures? Many NGO-based INSET projects in South Africa work within a paradigm which may be summarised as follows:

CAUSE	PROBLEM	SOLUTION	INDICATORS OF SUCCESS	RESULT
rote teaching methods, inappropriate curricula	poor quality education	change to child-centred teaching methods and problem-centred, activity-based materials	• group-based activities • pupil-pupil and pupil-teacher talk • teacher not prescribing methods	improved quality of education
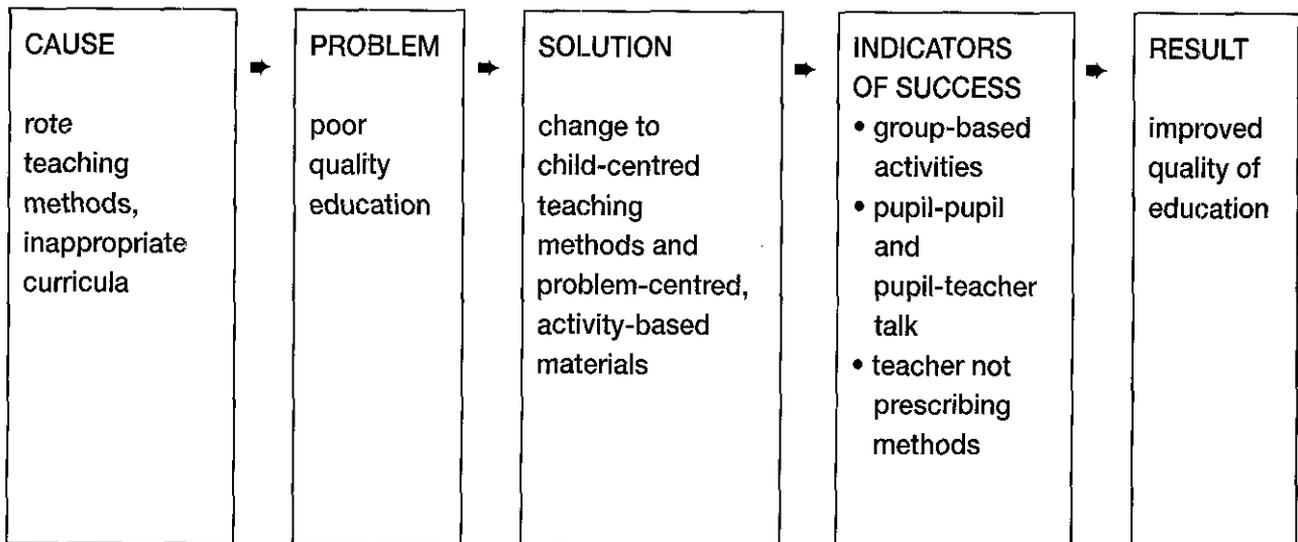
Schollar's evaluation of the Palabora INSET programme seriously undermines confidence in this paradigm in two ways:

◆ Success in changing what happens in the classroom depends to a considerable degree on factors beyond the control of the classroom teacher.

◆ The indicators of success listed above do not themselves reflect an improved quality of teaching or learning, but an intermediate state which may or may not lead to improved outputs. This confusion between an outcome and an intermediate state has been referred to as face validity (Cohen and Manion, 1980).

Both findings are corroborated by a number of other studies and we will return to these issues in section 3 below. The methodological point to make at this stage is that intermediate states should not be confused with outcomes indicative of success. But, identifying what is meant by quality and finding ways of measuring it are both vexed questions.

Weingart (Muller and Vinjevold, 1995) notes a move amongst evaluators internationally towards focusing on outcome indicators as the most reliable measure of programme impact. This trend is beginning to emerge in South Africa. Thus, for *Schollar*, the ultimate objective of an INSET programme is to impact on learning, "as measured by sustained enhancement of pupil performance." *(ibid, 27)*. *Dyrenfurth* supports this view to the exclusion of 'process' considerations, while *Khulisa* and *Bateson (1995)* both combine an examination of classroom practices with quantitative measures of learning outcomes.

Comparison. Once a suitable measure of success has been identified, assessing whether change has occurred requires a comparison. A surprising number of evaluators not only managed to discern change on the strength of one-off observations of lessons with a limited number of teachers, but also considered their method rigorous enough to attribute this change to the programme under evaluation (eg. *Adler, Dlamini*). Such findings cannot be accepted as valid. Discerning change and attributing it to a specific intervention are the most problematic

aspects of evaluation, requiring particular care in design, instrumentation and data testing. Controlled comparison is a sine qua non.

Comparison may be achieved through:

◆ A pre/post test method. None of the evaluations under review employed this technique, largely because of a shortage of time, and the fact that the programme was in progress by the time the evaluation was commissioned. Indeed, having the programme subjects available in a pre-intervention state is a very rare situation.

◆ An experimental/control group design. This is the most common form of comparison and was used by *Dyrenfurth, Bateson (1995)* and *Khulisa*.

◆ Experimental/baseline data design. No suitable baseline data exists in South Africa at present, and a number of evaluators have called for the construction of such measures.

◆ Expert judgement. While Weingart (Muller and Vinjevold, 1995) insists that quantification is essential for reliable comparison, we would not dismiss the use of unstructured expert judgement to identify at least tentative indications of change. *Schollar*, for example, noted that teachers who had been on the Palabora programme longer were more confident and adroit with interactive classroom methods than newcomers. This is a very important observation that casts a more optimistic light on what might otherwise be a rather pessimistic conclusion concerning the future of such practices. In general, however, comparisons are facilitated and conclusions concerning change more convincingly argued if methods involving expert judgement are structured and quantified.

2.6 Sifting and testing data
Triangulation. *Schollar* used `triangulation' to test his self report data; his classroom observations caused him to reject the views of teachers, principals and other stakeholders that the programme is improving the quality of teaching and learning.

Statistics. Quantitative data is best tested by statistical means: *Dyrenfurth* was able to ascertain that the differences in mean scores across his experimental and control groups were unlikely to be due to chance variations within each group, through an analysis of covariance.

Parts and the whole. Only once suitable indicators of success have been identified, a change in these indicators established, and the variables controlled in such a way that the change is attributable beyond reasonable doubt to a specific intervention, can it be said that the intervention is successful in the sample under evaluation. However, it still cannot be said that the programme as a whole is successful: in order to be able to reach this conclusion, the evaluator must demonstrate that the sample tested is representative of the programme population as a whole.

Only *Dyrenfurth* was able to meet this criterion, and then only because he was evaluating a pilot project and used the entire pilot as his sample. It remains to be seen whether the project can deliver the same results on a larger scale; the evaluator has doubts and includes recommendations on steps he considers necessary for impact on a larger scale.

Bateson (1995) involved the project under evaluation in a self-sampling technique in which six of SEP's ten regions were asked to pick ten 'typical' project schools, and ten control schools which matched the project schools on a number of criteria. While the sample is certainly large enough to satisfy the demands of statistical significance, the method of obtaining the sample must suffer from the same weaknesses as other forms of self report data. He also does not outline the criteria for choosing the six sampled regions. As a consequence, the relationship between the 60 schools included in the sample and the approximately 1200 schools which the project claims to service must be open to question.

In any event, as *Khulisa* notes, no defensible sampling process can be conducted given the incomplete state of the data bases of all the projects under evaluation. The existence of an accurate data base is a prerequisite for any systematic sampling procedure.

2.7 Reliability and validity

The reliability of a set of data or conclusions refers to the degree to which the findings are likely to be repeated, at a different time, with a different sample of the programme population or with a different evaluator (Cohen and Manion, 1980). Naturally different evaluators bring specific strengths and experiences to bear, and hence produce different insights. However, in respect of the evaluation per se - deciding whether the evaluatee has been successful in meeting its goals - the criterion of reliability demands that the overall conclusion will coincide with that of another independent study.

Validity refers to the degree to which the evaluation succeeds in measuring what it has set out to measure. For *Bateson (1995)* validity means:

> "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment"
>
> *(Messick, quoted in Bateson, 1995, 1)*

Our methodological discussion suggests three criteria for assessing the reliability and validity of any evaluation which attempts to assess the quality of any INSET programme:

◆ Because of the problem of intervening states, a comparative quantitative analysis of pupil learning gains is the most effective method of establishing whether an INSET programme has resulted in an improvement in the quality of schooling. It goes without saying that this component should follow the established rules for experimental design and statistical significances.

◆ A qualitative analysis, which relies heavily on expert classroom observation, is necessary to investigate the reasons for any pupil learning gains which may or may not

occur. In order to allow for alternative interpretations, all relevant evidence should be presented in support of the argument.

◆ The study as a whole should provide a systematic argument linking its conclusions to a description of the project and its intentions, the identification of outcomes appropriate to testing the successful realisation of these intentions, and the effects of the project as ascertained through the relevant quantitative and qualitative data sources.

There is no such thing as a perfect evaluation; nor is any evaluation without illuminative insights. In our discussion below of the findings of the 50 studies under review, all significant conclusions are drawn only from those studies, or parts thereof, considered to be more reliable.

3. FINDINGS OF THE EVALUATIONS

3.1 Boundary conditions
One of the most striking and widespread findings is the low work rate in many South African schools. A number of evaluators mentioned this as a factor which severely inhibited both their own work, and the successful implementation of the programme under evaluation. *Khulisa*, for example, demonstrates a strong correlation between overall management and the environmental quality of the school, and good test results amongst pupils.

Poor school management is a major factor in the majority of schools serviced by the projects covered by this review. However, it is clear that these practices have become part of institutional and community cultures. As such, their eradication which is only likely to occur at a relatively slow ameliorative rate - requires far more than INSET with individual teachers, with or without management training for principals and other leaders. Change will need to involve the entire school community - pupils, teachers, management, parents and provincial and district officials - and focus on institutional development.

The magnitude of the problem is indicated by *Bateson's (1994)* estimate that 50 days of the school year are lost to effective education; this is corroborated by *Peacock's* estimate that the average school week is reduced by around 40%. Specific practices giving rise to this situation include:

◆ an inordinate amount of time given to timetabling *(Schollar)* and testing *(Peacock)*;
◆ early closing of the school day *(Schollar)*, and the week *(Peacock, Bateson, 1994)*; no school on payday *(Bateson, 1994)*;
◆ lengthy preparations for athletics *(Peacock, 1995; Schollar, Khulisa)*;
◆ violence and strikes *(Jansen and Perold; Bateson, 1994)*
◆ doubling of classes *(Schollar, Peacock)*.

It seems probable that increasing the work rate of schools will have an immeasurably greater effect on the productivity and competitiveness of the country than any amount of INSET or management training on their own. *Schollar's* suggestion that making the school responsible for the establishment and maintenance of certain boundary conditions as a prerequisite for the implementation of INSET programmes in the school is only one of a number of institutional approaches being mooted from a variety of quarters. These initiatives are discussed in the main report of the NGO component of the Teacher Education Audit.

3.2 INSET model and project management
The overall conceptualisation of any INSET programme, and its strategic planning and management are closely related. Collectively, the evaluators identified five priority issues.

Tension between teachers' perceived needs and project priorities. In their attempts to develop innovative solutions to curricular and pedagogical problems, projects can run ahead of teachers' abilities to relate these innovations to the everyday demands of their classrooms, such as getting through the syllabus, coping with large classes and preparing their students for exams *(Adler, Mareka, Schollar)*. This problem is greatly exacerbated by the fact that many NGOs are antagonistic to the current curriculum, and their programmes are directed towards fundamentally restructuring it.

Coordination with government. Innovation and change always places those responsible for its implementation in a state of tension. But the tension noted above would be relieved if INSET initiatives directed towards classroom innovation were conceived within a broader programme of curriculum reform, which is at least approved by government if not initiated and drawn by provincial and district authorities.

In-school support of teachers is widely regarded as important in improving the chances of innovative practices being taken up by teachers *(Adler, Khulisa, Mareka, Jansen and Perold*, and *Wederkind)*. However, this is a labour intensive activity and may require long periods of sustaining before effects are noticed *(Adler)*. It is thus extremely costly. Some programmes, notably SEP *(Bateson, 1995)* and PMP *(Khulisa)*, are addressing this issue through models in which government subject advisors, inspectors or seconded teachers take responsibility for specified project activities. Indeed, the third and final stage of the PMP model envisages government taking responsibility for the programme and implementing it in all schools.

Management information systems (MIS). Three evaluators *(Adler, Schollar and Khulisa)* noted that their work was greatly hampered by the absence of a proper MIS on the part of the evaluatee. The main report of the NGO indicates that this problem is almost universal amongst INSET NGOs. *Khulisa* makes the point that without an MIS the project is unable to assess its own qualitative impact and scale of activity, and hence unable to calculate unit costs.

Research. A number of NGOs are ardent advocates of their own work. Their arguments are driven by the construction of normative models based on untested assumptions, rather than by on-going assessment of their achievements. They have leapt to Korten's (see section 1) third stage - policy advocacy - without consolidating the first two. There is a high degree of consensus amongst evaluators *(Bateson, 1995; Dyrenfurth; Adler; Khulisa, Schollar)* that projects need to continuously research and evaluate the efficiency and impact of their models.

Certification. The absence of suitable incentives for teachers to attend and implement INSET programmes is identified as a common problem, and many evaluators recommend certification of the programme as a solution. *(Schollar, Khulisa, Adler, Peacock)*.

For *Schollar*, certification should include an assessment of competence.

3.3 Effects on teachers

According to many of the evaluations under review it is easier to change the form of teaching practices than their substance *(Adler, Schollar, Khulisa, Peacock)*.

> "The struggle to introduce an alternative and more enlightened methodology has been won. The next target is becoming clear: to re-introduce and reinforce awareness of the need for more cognitive effort to be routinely required of pupils. It is not nearly enough for teachers to merely 'interact' with pupils, without reference to the quality of the work being demanded from the pupils."
>
> *(Peacock, 2)*

Khulisa noted a correlation between an improvement in teaching quality and:

♦ smaller class sizes
♦ length of time of teachers on INSET programme
♦ increasing age of teachers.

Schollar points out that improved teaching quality does not involve:

> "the simple application of a supposedly 'alternative' or 'progressive' methodology, but the **extension of the capacity of teachers to use appropriate methodologies:** appropriate to the topic, the class and the school."
>
> *(op cit, 17)*

These observations give support to the reservations we expressed in section 2.5 above against using the forms of teacher behaviour as measures of INSET success.

3.4 Effects on pupils

Evaluators are increasingly recognising the importance of focusing on pupil performance as the most valid indicator of INSET success *(Adler, Schollar, Jansen and Perold, Khulisa, Peacock)*. In this respect, *Peacock* presents a sobering picture of many interactive classrooms:

> "most of the children most of the time do nothing... and certainly nothing as cognitively demanding as individual reading and writing"
>
> *(ibid, 13)*

Khulisa found that, while project children were able to communicate their solution strategies to arithmetic tasks better than their control group counterparts, they did not score a significantly higher number of correct answers. An examination of the solutions employed by the project children raises the disturbing possibility of constructivist methods actually being counterproductive : by trapping children in ad hoc, 'baby' methods of calculating they are prevented from acquiring a systemic perspective of the discipline of mathematics and higher order algorithmic processes. These arguments clearly illustrate the dangers of INSET projects being driven by evangelical zeal for a particular method rather than by a research-based, enquiring approach.

However, significant rays of light are provided by the *Dyrenfurth* and *Bateson (1995)* studies demonstrating positive pupil learning gains as a result of the PROTEC and SEP projects.

3.5 Costs

The establishment of accurate MIS processes, and separate cost centres for different project activities are prerequisites for calculating the costs of INSET work *(Schollar, Khulisa)*.

Adler notes that the high cost and relatively small returns on activities such as classroom-based materials development, and in-school teacher support do not warrant their continuation.

Khulisa advocates that a minimum of 25% of costs should be recovered from teachers, although this is only feasible if the courses are certified

4. CONCLUSION

4.1 INSET innovation and evaluation

NGOs involved in INSET are concerned with finding innovative solutions to the problem of the poor quality of the schooling system. Evaluation and innovation go hand in glove. Without continuous assessment we have no way of valuing these INSET initiatives: whether to expand, close down, fine-tune or redirect. Making decisions about allocating resources is like shooting in the dark. South African evaluation practice is in need of a great deal of muscle toning. Our analysis concludes that, in order to assess programme quality, evaluations should contain three essential components:

◆ A systematic and well substantiated argument linking project intentions, activities and outcomes.
◆ A quantitative assessment of suitable project outcomes - preferably pupil learning gains - controlled by means of appropriate comparative techniques.
◆ A qualitative assessment - which should probably include careful classroom observation - into the reasons for any quantitative changes detected.

4.2 Evaluation review questions

Our conclusions are shaped by three questions:

Q1a: Are NGOs able to develop and implement new models of INSET which impact on schools and classrooms so as to make teaching and learning more meaningful and effective?

Answer: A qualified yes.

Q1b: What conditions would optimise such impact?

Answer:
◆ Establishment of an accurate management information system.
◆ Explicit description of the INSET model and its assumptions.
◆ Continuous research and evaluation of project outcomes.

Q2a: Are NGOs able to provide INSET on the kind of scale required to improve the quality of substantial sectors of the schooling system to any significant degree?

Answer: Perhaps

Q2b: What conditions would facilitate NGO delivery to scale?

Answer:
◆ An unqualified positive answer to Q1a.
◆ Working closely with government within a clearly articulated division of labour.
◆ Certification, which includes an assessment of classroom competence.

Q3: What conditions would promote NGO INSET delivery in an affordable manner?

Answer:
◆ A careful analysis of component costs, through the establishment of cost centres for different programme activities.
◆ Assessment of the affordability of the various components, by comparing costs with educational returns.
◆ Inventing ways of recovering costs through the sale of services to teachers and government departments; through RDP grants; and from donors.

BIBLIOGRAPHY

Cohen, L and Manion, L (1980) *Research Methods in Education* (London, Croom Helm)

IDT (1995) *Survey of the Financial State of 128 South African Non-Governmental Organisations (NGOs) and Community-Based Organisations (CBOs).*

Korten, DC (1990) *Getting to the 21st Century.* (West Hartford, Kumarian Press)

Muller, JP and Vinjevold, PA (1995) *Weingart on Quality Assessment : What it Means for Project Evaluation* (Report on a seminar led by Peter Weingart, Stellenbosch)

Leat, D (1990) Voluntary Organisations and Accountability : Theory and Practice, in HK Anheier and W Seibel (eds.) *The Third Sector : Comparative Studies of Non-profit Organisations.* (New York, Walter de Gruyter)

Weiss, CH (1990), in MC Alkin (ed.) *Debates on Evaluation* (Newbury Park, Sage Publications)

Adler, J (1995) *Looking Back, Inside and Ahead :* An Evaluation of the Mathematics Education Project (MEP).

Angelis, D, Shongwe, S and Welch, T. (1995) *Evaluation of ORT-STEP's Teacher Training Programme.*

Angelis, D. (1994) *An Evaluation of PROTEC's Tutor Training Programme.*

Bahr, MA and Perold, H. (1993) *Report of the evaluation of the Spider's Place comic as an educational resource in a rural schooling environment.*

Bateson, DJ. (1994) *Changing Times : An Evaluation of CASME*

Bateson, DJ. (1995) *Science and Teachers : An Evaluation of the Science Education Project (SEP).*

Bell, DI. (1995) *An investigation into the Influence of an INSET Programme on Teacher Cognitions.*

Budlender, D. (1994) *Evaluation of the Teaching and Learning Resources Centre.*

Campbell, D. (1993) *The Teaching and Learning Resources Centre : Report on Courses.*

Cliff, A. (1994) *Independent Evaluation report on the 1994 Novalis Teacher Enrichment Programme.*

Connolly, E. (1992) *The Edu-Train Project : The Reduction of Prejudice.*

de Jong, T. (1994) *Composite Evaluation Report: Teacher Inservice Project.*

Dlamini, T. (1994) *An Evaluation of the Maths Centre for Primary Teachers (MCPT).*

Dyrenfurth, MJ. (1995) *RSA PROTEC Technology Education Pilot : Initial Evaluation Report.*

Ferrandi, NCH. (1993) *Evaluation of the Science and Mathematics Subject Support Programme.*

Gains, P. (1995) *An Evaluation of the SMILE Programme.*

Gilmour, JD and Soudien, CA. (1994) *An Evaluation of the Management Component of the TOPS Programme.*

Gunter, A. (1994) *Report on the Master Maths involvement with In-Service Teacher Training in Mathematics in 1993 in the Western Cape.*

Harpring, SA. (1993) *Evaluation of the In-Service Education of Teachers (INSET) Programme.*

Hartshorne, K. (1994) *Evaluation of the English Language Education Trust, Durban.*

Hartshorne, KB. (1992) Brief Review of : Schools English Language Programme (SELP) and Teachers English Language Improvement Project (TELIP) : *Origins, Development, Present Status and Future Potential*

Henning, E. (1994) *Evaluation of the Education Support Project (ESP).*

Institute for International Research *Improving Educational Quality Project* : An Evaluation of the TOPS Management Programme.

James, CN (1994) *Thlabane Investigative Maths Scheme.*

Jansen, J and Perold, H. (1994) *How do Teachers Use Innovative Primary Science Materials*: Report on an Evaluation of Spider's Place.

Johanson, EK. (1993) *Evaluation of the Secondary Schools English Research Project (SSERP).*

Khulisa Management Services cc. (1995) *Evaluation of the South Africa Primary Mathematics Project : Centre for Productive Education Consortium.*

Kroes, H. (1991) *The Evaluation of the "Easing into English Course"*

Kroes, H. (1993) *Evaluation of the Reading and Cognition Project (START).*

Lategan, A. (1991) *Evaluation - A Series of Reports on the Development of the Programme in the Eastern Cape, Natal, Orange Free State, Transvaal and the Western Cape, and a Composite Report on the Development of the Programme Nationally.*

Levy, S and Mphahlele, M. (1994) *Evaluation of the Count Project in the Eastern Cape Region.*

Malone, JA (1994) *Impressions of the Rhodes University Mathematics Education Project : Certificate in Primary Mathematics Education.*

Mareka, CL. (1992) *Evaluation of School Management Programme : University of Zululand.*

McKay, V. (1993) *An Exploration of Pupils' Perceptions of Their Teachers with Special Reference to Teachers Who Have Received Training on the Skills Programme* : Research Project undertaken for the Centre for Cognitive Development.

Mercorio, G and Morris H. (1994) *An Assessment of the Young People's Theatre Educational Trust.*

Moonsammy, G, van der Westhuizen, GJ and Abrahams, M. (1994) *Evaluation of the CCD INSET Programme.*

Moonsammy, JG and Mononyane, T. (1994) *Evaluation of the CCD Programme in Mamelodi Primary Schools.*

Muller, J. (1990) *Evaluation of the Teachers Learning and Resource Centre,* University of Cape Town.

Nell, M and Shapiro, JB. (1995) *Report on St Mary's DSG Outreach Programme.*

Nickson, M and Julie, C. (1994) *The Primary Mathematics Project.*

Peacock, M. (1995) *Evaluation of the Natal College of Education DE2 Distance Learning Course For Primary School Teachers.*

Peacock, M. (1995) *Natal College of Education DE2 Distance Learning Course for Primary School Teachers* : Summary of issues arising from the Evaluation.

Pratt, EE. (1992) *Independent Evaluation Report on the Centre for Cognitive Development.*

Robinson, M. (1994) *Evaluation of the Centre for Cognitive Development.*

Schollar, E and Associates cc. (1995) *Evaluation of the Teacher Upgrading Project of the Palabora Foundation.*

Serote, CAK. (1994) *Evaluation of the Lenyenye INSET Project.*

Stadler, C. (1994) *Evaluation of the Education Support Project (ESP).*

Tselane, MB. (1994) *Evaluation of the Community Upliftment through Education Programme.*

Ural, I and Ratele, K. (1994) *Evaluation of CAI In-Service Computer Programmes in Mathematics at the Siza Centre.*

Volmink, J and Hardman, S. (1992) *The Maths Centre for Primary Teachers : A Case Study in Evaluation Management.*

Volmink, JD. (1992) *Centre for Cognitive Development,* Vista University.

Volmink, JD. (1993) *Evaluation of the Maths Centre for Primary Teachers.*

Wedekind, V. (1995) *Evaluation of the MEDU Active Classroom Learning Programme.*

Zabala, G and Moja, T. (1994) *Evaluation of the Education Support Project (ESP).*

DELEGATES

Adam, Fatima
J E T
PO BOX 178, WITS, 2050
tel: (011) 403-6401, fax: (011) 339-7844

Adler, Jill
University of Witwatersrand - Education Dept
P/Bag 3, Wits, 2050
tel: (011) 716-5257, fax: (011) 339-3956

Ainslee, Susan
Midlands Education Development Unit (MEDU)
P/Bag X01, Scottsville, 3209
tel: (0331) 260-5365, fax: (0331) 260-5080

Barry, Noleen
Independent Training & Educational Centre (ITEC)
PO Box 549, East London, 5200
tel: (0431) 43-8333, fax: (0431) 29263

Basson, Ray
University of Witwatersrand - Education Dept
P/Bag 3, Wits, 2050
tel: (011) 716-5222, fax: (011) 339-3956

Beck, Carol
Young Peoples' Theatre Educational Trust
PO Box 472, Athlone, 7760
tel: (021) 697-1881, fax: (021) 697-1881

Bisgard, Jennifer
Khulisa
PO Box 26992, Hout Bay, 7800
tel: (021) 790-3831, fax: (021) 790-5407

Bloch, Graeme
J E T
PO Box 178, Wits, 2050
tel: (011) 403-6401, fax: (011) 339-7844

Browne, Daniela
St Mary's Interactive Learning Experience (SMIILE)
PO Box 47158, Greyville, 4023
tel: (031) 764-6336, fax: (031) 764-6336

Burmeister, John
Molteno Project
PO Box 30696, Braamfontein, 2017
tel: (011) 339-6603, fax: (011) 339-3555

Campbell, Ernie
PROMAT
PO Box 27265, Sunnyside, 0132
tel: (012) 343-2275, fax: (012) 344-5844

Comrie, H
Natal College of Education
P/Bag 9007, Pietermaritzburg, 3200
tel: (0331) 45-4515, fax: (0331) 945-5303

Coombe, Jane
Mathematics Education Project (MEP)
University of Cape Town, P/Bag Rondebosch, 7700
tel: (021) 650-3368, fax: (021) 650-3489

Cooper, Charlie
English Language Teaching Information Centre
(ELTIC)
PO Box 32398, Braamfontein, 2017
tel: (011) 403-3238, fax: (011) 403-3240

Currin, Sally
St. Mary's DSG Outreach Project
PO Box 11379, Hatfield, 0028
tel: (012) 342-3189, fax: (012) 342-3189

Dawjee, Rabia
Teachers' English Language Improvement Project (TELIP)
P/Bag X3, Wits, 2050
tel: (011) 716-5489, fax: (011) 339-7763

Delport, Deon
Wits Rural Facility
P/Bag X420, Acornhoek, 1360
tel: (01528) 33991, fax: (01528) 33992

Diphofa, Mashwahle
J E T
PO Box 178, Wits, 2050
tel: (011) 403-6401, fax: (011) 339-7844

Dooley, Bridget
JCI
PO Box 590, Johannesburg,2000
tel: (011) 373-2802, fax: (011) 492-1070

Dovey, Ken
Institute for Social and Individual Development in
Africa (ISIDA)
PO Box 7426, East London, 5200
tel: (0431) 22539, fax: (0431) 438307

Dube, Syndney
Institute of Education & Human Development
P/Bag X1001, KwaDlangezwa, 3886
tel: (0351) 93243, fax: (0351) 93243

Du Plessis, Lind
Optima College of SANCB
PO Box 11144, Brooklyn, 0011
tel: (012) 346-1190, fax: (012) 346-1149

Dwight, Mike
Read Educational Trust (READ)
PO Box 30994, Braamfontein, 2017
tel: (011) 339-5941, fax: (011) 403-2311

Eisenberg, E
Ort-Step Project
P/Bag X13, Halfway House, 1685
tel: (011) 651-6536, fax: (011) 651-6428

Erwee, Jan
Centre for Continuing Education (CENCE) UPE
PO Box 1600, Port Elizabeth, 6000
tel: (041) 55-8718, fax: (041) 56-4519

Flanagan, Wendy
PREP at UCT
University of Cape Town, P/Bag Rondebosch, 7700
tel: (021) 650-9111, fax: (021) 650-3489

Gardiner, Michael
Education Support Project (ESP)
PO Box 31511, Braamfontein, 2017
tel: (011) 339-2389, fax: (011) 339-3334

Glover, Peter
Primary Science Programme (PSP)
P O Box 32198, Braamfontein, 2017
(011) 403-2696, (011) 339-2962

Griessel, Hanlie
University of Natal - Teach Test Teach
P/Bag X10 Dalebridge 4014
tel: (031) 260-3259, fax: (031) 260-3266

Hassett, Alex
St. Mary's DSG Outreach Project
PO Box 11379, Hatfield, 0028
tel: (012) 342-3189, fax: (012) 342-3189

Hollingworth, Barbara
Secondary Schools English Research Project (SSERP)
PO Box 93441, Yeoville, 2143
tel: (011) 484-7774, fax: (011) 484-3121

Jansen, Jonathan
University of Durban-Westville - Dept of Curriculum
Studies
P/Bag X54001, Durban, 4000
tel: (031) 820-2255, fax: (031) 820-2866

John, Beulah
Language in Learning & Teaching Project (LILT)
P/Bag X01, Scottsville, 3209
tel: (0331) 260-5847, fax: (031) 260-5575

Joubert, Betsy
Project for the Establishment of Preprimary and Primary
Schools
PO Box 962, Irene, 1675
tel: (012) 667-1016, fax: (012) 667-1202

Khomo, Preston
Palabora Foundation
PO Box 1263, Phalaborwa, 1390
tel: (01524) 2546, fax: (01524) 85442

Kruger, Andrew
National Language Project (NLP)
PO Box 378, Salt River, 7924
tel: (021) 47-2760, fax: (021) 47-2994

Kwazi, Khethiwe
JCI
P O Box 590, Johannesburg, 2000
tel: (011) 373-2626, fax: (011) 492-1070

Louw, Johann
University of Cape Town - Dept of Psychology
P/Bag, Rondebosch, 7700
tel: (021) 650-3414, fax: (021) 650-3657

Makume, TI
Teacher Opportunity Programmes (TOPS)
PO Box 428, Kroonstad, 9500
tel: (0562) 3723, fax: (0562) 51299

Mbelle, Zo
J E T
PO Box 178, Wits, 2050
tel: (011) 403-6401, fax: (011) 339-7844

McLean, Hugh
J E T
PO Box 178, Wits, 2050
tel: (011) 403-6401, fax: (011) 339-7844

McRobert, Marianne
Nelson Mandela Children's Fund
33 Scott St, Waverley, 2090
tel: (011) 786-9140, fax: (011) 786-9197

Meyer, Susan
National Business Initiative
PO Box 294, Auckland Park, 2006
tel: (011) 482-5100, fax: (011) 482-5507

Milford, H P
Namaqualand Resource and Educational Centre
(NAMREC)
PO Box 502, Springbok, 8240
tel: (0251) 22726, fax: (0251) 22726

Modiba, Maropeng
University of Witwatersrand - Education Dept
P/Bag 3, Wits, 2050
tel: (011) 716-5820, fax: (011) 339-3956

Moodley, Juliann
Development Evaluation Unit DBSA
614 Drieankerhof, 214 Troye St, Sunnyside, Pretoria, 0002
tel: (012) 44-6488

Moolla, Nadeen
Teacher In-Service Project (TIP)
P/Bag X17, Bellville, 7535
tel: (021) 959-2911, fax: (021) 959-2647

Morake, Dorcas
Education Dept, University of Botswana
Gaborone, Botswana
fax: (09267) 35-2888

Morphet, Tony
University of Cape Town, Dept of Adult Education
P/Bag, Rondebosch, 7700
tel: (021) 650-3252, fax: (021) 650-2893

Morris, Gill
St Andrews Outreach Foundation,
Teacher Enrichment Programme
PO Box 79001, Senderwood, 2145
tel: (011) 453-9408, fax: (011) 453-1112

Muller, Joe
University of Cape Town - Education Faculty
P/Bag, Rondebosch, 7700
tel: (021) 650-2770, fax: (021) 650-3489

Naidoo, Bunny
Primary Maths Project (KwaZulu-Natal)
100 Davenport Sq, 89 Davenport
tel: (031) 21-4244, fax: (031) 21-9011

Ntombela, G M
Centre for the Advancement of Science and Maths
Education (CASME)
PO Box 17112, Congella, 4013
tel: (031) 81-1390, fax: (031) 81-5070

Pease, Jean
Primary Open Learning Pathway Trust (POLP)
Rycom Cntre, 2nd Fl, c/o Old Klipfontein/Murton Rd,
7764
tel: (021) 637-8094/5, fax: (021) 638-4410

Pelser, Ann
English Language Education Trust (ELET)
6th Fl, 74 Aliwal Str, Durban,4001
tel: (031) 32-0501, fax: (031) 37-0002

Persens, J
Teacher Advancement in Maths - UWC
P/Bag X17, Bellville, 7535
tel: (021) 959-3059, fax: (021) 959-3055

Prinsloo, Esme
Careers Research & Information Centre (CRIC)
PO Box 378, Athlone, 7760
tel: (021) 637-8040, fax: (021) 633-9022

Pullen, Dee
Boitumelo Early Learning Centre
PO Box 56348, Arcadia, 0007
tel: (012) 326-0484, fax: (012) 326-0484

Rehman, Ayesha
J E T
PO Box 178, Wits, 2050
tel: (011) 403-6401, fax: (011) 339-7844

Reid, Kate
Segal Quince Wicksteed Limited
PO Box 17, Faerie Glen, Pretoria, 0043
tel: (012) 991-2445

Roberts, George
Developing Education for Tomorrow Trust (DEVED)
PO Box 121, Bramley, 2018
tel: (011) 887-9488, fax: (011) 887-9488

Rossouw, Lynn
Primary Maths Projects (W. Cape)
P/Bag X17, Bellville, 7530
tel: (021) 959-2510, fax: (021) 951-2602

Salmon, Cecily
Fulcrum Teacher Education Project
PO Box 1608, Durban, 4000
tel: (031) 21-4244, fax: (031) 21-9011

Scott, B
Winterberg Schools Trust
PO Box 123, Tarkastad, 5370
tel: (04582) 4130,

Sehoole, Trevor
J E T
PO Box 178, Wits, 2050
tel: (011) 403-6401, fax: (011) 339-7844

Shan, Sharnjreet
Maths Centre for Primary Teachers (MCTP)
PO Box 117, Auckland Park, 2006
tel: (011) 726-4237/8, fax: (011) 726-1510

Silva, Peter
Sakhisizwe Inset Unit
PO Box 304, Grahamstown, 6140
tel: (0461) 27115, fax: (0461) 24457

Smit, J H
Institute of Maths and Science Teaching (IMSTUS)
University of Stellenbosch
P/Bag X1, Matieland, 7602
tel: (021) 808-3484, fax: (021) 808-3000

Smith, Penny
Cooperative Organisation for Upgrading of Numeracy
Training (COUNT)
PO Box 55072, Northlands, 2116
tel: (011) 339-7376, fax: (011) 339-7376

Smith, Leonard
Novalis Teacher Enrichment Programme
PO Box 53090, Kenilworth, 7745
tel: (021) 797-1857, fax: (021) 761-0057

Steyl, Elize
R A U - Faculty of Education
PO Box 524, Auckland Park, 2006
tel: (011) 489-2911, fax: (011) 489-2781

Sullivan, Pat
Management of Schools Training Programme (MSTP)
PO Box 633, Auckland Park, 2006
tel: (011) 482-1782, fax: (011) 482-1786

Taylor, Nick
J E T
PO Box 178, Wits, 2050
tel: (011) 403-6401, fax: (011) 339-7844

Tema, Botlhale
Programme for Technology Careers (PROTEC)
PO Box 52657, Saxonwold, 2132
tel: (011) 788-7209, fax: (011) 880-1795

Tiernan, Brigid
Catholic Institute of Education (CIE)
PO Box 2083, Southdale, 2135
tel: (011) 433-1888, fax: (011) 680-9628

Tracey, Kathy
J E T
PO Box 178, Wits, 2050
tel: (011) 403-6401, fax: (011) 339-7844

Valla, Vijay
J E T
PO Box 178, Wits, 2050
tel: (011) 403-6401, fax: (011) 339-7844

van der Linde, Jack
Research Institute for Education Planning (RIEP)
PO Box 339, Bloemfontein, 9300
tel: (051) 401-2856, fax: (051) 47-4939

van der Westhuizen, Gert
Centre for Cognitive Development (CCD)
PO Box 8385, Hennopsmeer, 0046
tel: (012) 663-1204, fax: (012) 663-1321

van Niekerk, Saar
Tecknikon SA
P/Bag X6, Florida, 1710
tel: (011) 471-2000, fax: (011) 471-2134

Vilakazi, J
Midlands Educational Trust
PO Box 80, Howick, 3290
tel: (0332) 36-6993, fax: (0332) 30-6993

Vinjevold, Penny
J E T
PO Box 178, Wits, 2050
tel: (011) 403-6401, fax: (011) 339-7844

Weingart, Peter
Bielefeld University, Germany
e-mail: pweingart@bird.zif.uni.bielefeld.de

White, Helen
Rhodes Univ. Mathematics Education Project (RUMEP)
c/o Education Dept, Rhodes University, Grahamstown, 6140
tel: (0461) 31-8166, fax: (0461) 25642